2013

Computational prediction of RNA-protein interaction partners and interfaces

Usha Muppirala
Iowa State University

Computational prediction of RNA-protein interaction partners and interfaces

by

Usha Kiran Muppirala

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Drena Dobbs, Co-Major Professor
Robert Jernigan, Co-Major Professor
Heike Hofmann
Guang Song
Edward Yu

Iowa State University

Ames, Iowa

2013

TABLE OF CONTENTS

ABSTRACT

RNA-protein interactions play important roles in fundamental cellular processes involved in human diseases, viral replication and defense against pathogens in plants, animals and microbes. However, the detailed recognition mechanisms underlying these interactions are poorly understood. To gain a better understanding of the molecular recognition code for RNA-protein interactions, this dissertation has three related goals: i) to develop methods for predicting RNA-protein interaction partners; ii) to develop an approach for predicting interfacial residues in both the RNA and protein components of RNA-protein complexes; and iii) to develop computational tools and resources for investigating RNA-protein interactions.

First, we present machine learning classifiers for predicting RNA-protein interaction partners. The classifiers use the amino acid composition of proteins and the ribonucleotide composition of RNAs as input to predict whether a given RNA-protein pair interacts. We show that protein and RNA sequences alone (i.e., in the absence of any structural information) contain enough signal to allow reliable prediction of interaction partners.

Second, we present RPISeq, a webserver that predicts the interaction probabilities of input RNA-protein pairs, using the above-mentioned machine learning classifiers. A comprehensive database of RNA-protein interactions, RPIntDB, is integrated with the webserver to allow users to search for homologous proteins and their known interacting RNA partners.

Finally, we perform an analysis of contiguous interfacial amino acids and ribonucleotides in RNA-protein complexes for which structures are known. We generate a

dataset of bipartite RNA-protein motifs that can be used to predict interfacial residues in both the RNA and protein sequences of a given RNA-protein pair simultaneously. We show that taking binding partner information into account leads to higher precision in the prediction of RNA-binding residues in proteins.

Taken together, these studies have increased our understanding of how RNA and proteins interact.

CHAPTER 1. GENERAL INTRODUCTION

RNAs interact with proteins to regulate numerous cellular processes, ranging from DNA replication and transcription, to alternative splicing and translation (Hogan *et al.*, 2008, Licatalosi *et al.*, 2010). RNA-protein interactions (RPIs) also play important roles in human health and diseases, viral replication and pathogen resistance in plants (Kim *et al., 2009,* Sola *et al.*, 2011, Barkan, 2009). Still, a lot of questions need to be addressed related to the specificity and the mechanism of the underlying interactions between a protein and an RNA molecule.

The motivating questions behind this dissertation are: “How does a protein bind certain specific RNAs but not all RNAs? How do RNAs interact with specific proteins? What is responsible for this specificity?” Many computational studies on RNA-protein interactions have focused on small interfacial regions of RNA-protein complexes to understand specificity (Puton *et al.*, 2012). We refer to the problem of identifying interfacial residues as “interface prediction problem”. When interfaces are predicted on the protein (or RNA) molecule without considering interacting RNA (or protein) information, it is considered “non-partner-specific” prediction. These methods always predict the same set of interfacial residues even if the protein binds to different RNAs, using different subsets of those residues. When a method predicts interface residues on one molecule by considering the interacting partner molecule, it is termed “partner-specific” interface prediction. These prediction problems are different from “partner prediction problem”, which is a prediction of RNA interaction partner(s) for a known RNA binding protein, or protein binding partner(s) for an RNA. The starting datasets required for these types of computational methods are obtained

from the Protein Data Bank (PDB) (Berman *et al.*, 2000). Because experimental determination of RNA-protein complexes is difficult and time consuming, less than 2% of structures in the PDB are RNA-protein complexes. At the time this dissertation was initiated, no study had been published addressing the RNA-protein "partner prediction problem" or "partner specific interface prediction problem". Over the past 3 years, five papers have been published that describe computational methods to predict whether a given RNA and protein pair interacts (Pancaldi & Bähler, 2011, Bellucci *et al.*, 2011, Muppirala *et al.*, 2011, Wang *et al.*, 2013, Lu *et al.*, 2013). Also, high throughput experiments have begun to identify and characterize pairs of RNAs and proteins that participate in RPIs *in vivo* (Keene *et al.*, 2006, Licatalosi *et al.*, 2008, Ray *et al.*, 2009). This indicates the growing need for computational methods to predict RNA-protein interactions on a large scale. These methods are reviewed in detail in Chapter 2 of this dissertation.

Overall Goals

The overall research goal of this dissertation is to understand the determinants of molecular recognition in RNA-protein interactions and to identify features that can be used to accurately predict interaction partners and interfacial residues. My strategy has been to exploit available data from structure databases such as the PDB and sequence databases such as NPInter (Wu *et al.*, 2006) to develop computational tools for investigating and predicting RNA-protein interactions. Towards this goal, the following specific aims have been accomplished:

1. Develop a method to predict partners in RNA-protein interactions and demonstrate the application of this method to predict interactions in RNA-protein interaction networks (Chapter 3)
2. Develop RPISeq, a web server for predicting RNA-protein interaction partners (Chapter 4)
3. Develop RPIntDB, a comprehensive database of known RNA-protein interactions, to be used in conjunction with RPISeq (Chapter 4)
4. Analyze RNA-protein sequence motifs and develop a motif-based method to predict interfacial residues in RNA-protein complexes (Chapter 5)

In addition to the above specific aims, this dissertation also includes

5. An invited (peer-reviewed) summary of computational tools developed to date to investigate RNA-protein interactions (Chapter 2)
6. Back-end code for the RPISeq webserver and RPIntDB (Appendix 1)
7. A manuscript describing recent updates to the PRIDB, a Protein-RNA Interface Database (Lewis *et al.*, 2011) (Appendix 2)

Dissertation Organization

Chapter 1 is a brief overview of the work described in this dissertation.

Chapter 2 is a review paper published in the Journal of Computer Science and Systems Biology in 2013, entitled "*Computational tools for investigating RNA-protein interaction partners*". This invited peer-reviewed article discusses state-of-the-art methods available to predict partner specific RNA-protein interactions. It also summarizes the existing webservers and databases devoted to RNA-protein interactions. I conceived the study,

prepared the initial draft of the manuscript and participated in revisions and editing. Benjamin Lewis contributed to the discussion and editing. Drena Dobbs contributed to the study and revised the manuscript.

Chapter 3 is a research paper published in BMC Bioinformatics in 2011, entitled "*Predicting RNA-protein interactions using only sequence information*". This paper describes a new sequence based method to predict partner-specific RPIs using machine learning classifiers. I conceived the study, created the datasets, carried out the experiments, and prepared the initial draft of the manuscript. Drena Dobbs and Vasant Honavar contributed to the experimental design, supervised the work, and edited the manuscript.

Chapter 4 is a paper to be submitted to Bioinformatics, entitled "*RPISEQ & RPIntDB: Tools for predicting RNA-protein interactions"*. It describes RPIntDB, a database of RNA-protein interactions, and RPISeq, a webserver for predicting partner specific RNA-protein interactions. RPIntDB is a comprehensive collection of known RPIs extracted from the PDB, NPInter database and published high throughput experiments. A protein sequence of interest can be BLASTed against RPIntDB to identify homologous proteins and their known interacting RNA partners. Given a protein sequence(s) and RNA sequence(s), the RPISeq server predicts the probability of interaction between the input protein(s) and input RNA(s). I developed the webserver, implemented the database and prepared the manuscript. Drena Dobbs revised the manuscript.

Chapter 5 describes a novel method for predicting RNA binding residues in proteins and protein binding ribonucleotides in RNAs. This work entitled, "*A motif-based method for predicting interfacial residues in both the RNA and protein components of protein-RNA complexes*" is a manuscript in preparation. This study utilizes interfaces (i.e. derived from the

PRIDB) to create bipartite interfacial motifs. Given a pair of protein and RNA sequences, these motifs are then used as a guide to search for interface residues. Benjamin Lewis and I are co-first authors and contributed equally to the experimental design and manuscript preparation. Benjamin Lewis generated the motifs and contributed to data analysis. I conceived the algorithm and implemented the prediction method. Drena Dobbs contributed to the experimental design, supervised the work and edited the manuscript.

Chapter 6 summarizes the general conclusions of the dissertation, its potential applications and impacts. Future directions to extend this work are presented briefly.

Appendix 1 is a detailed description of the implementation of the RPISeq webserver. It includes pseudocode for the algorithm and organization of the server. The schema for RPIntDB and documentation useful for future updates are also recorded.

Appendix 2 is a database paper under revision for submission to Journal of Databases. It describes recent updates of the Protein-RNA Interface Database (PRIDB) developed by Benjamin Lewis. I contributed to the preparation of the manuscript, testing the functionality of the database manuscript revisions and editing.

References

Barkan A: Genome-wide analysis of RNA-protein interactions in plants. Methods Mol Biol 2009, 553:13-37.

Bellucci M, Agostini F, Masin M, Tartaglia GG: Predicting protein associations with long noncoding RNAs. Nature Methods 2011, 8:444-445.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. Nucleic Acids Res 2000, 28:235-42.

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO: Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol 2008, 6:e255.

Keene JD, Komisarow JM, Friedersdorf MB: RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nature protoc 2006, 1:302-7.

Kim MY, Hur J, Jeong S: Emerging roles of RNA and RNA-binding protein network in cancer cells. BMB Rep 2009, 42:125-130.

Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D: PRIDB: a Protein-RNA Interface Database. *Nucleic Acids Res* 2011, 39:D277-82.

Licatalosi DD, Darnell RB: RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 2010, 11:75-87.

Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer, Blume JE, Wang X, Darnell JC, Darnell RB: HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 2008, 456:464-9.

Lu Q, Ren S, Lu M, Zhang Y, Zhu D, Zhang X, Li T: Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics 2013, 14:651.

Muppirala UK, Honavar V, Dobbs D: Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics* 2011, 12:489.

Pancaldi V, Bähler J: In silico characterization and prediction of global protein-mRNA interactions in yeast. Nucleic Acids Res 2011:1-11.

Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM: Computational methods for prediction of protein-RNA interactions. J Struct Biol 2012, 179(3):261-8.

Ray D, Kazan H, Chan ET, Castillo LP, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR:. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nature Biotechnol 2009, 27:667-70.

Sola I, Mateos-Gomez PA, Almazan F, Zuñiga S, Enjuanes L: RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. RNA Biol 2011, 8:237-248.

Wang Y, Chen X, Liu Z-P, Huang Q, Wang Y, Xu D, Zhang XS, Chen R, Chen L. De novo prediction of RNA-protein interactions from sequence information. Mol BioSyst 2013, 9:133–42.

Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, Zhang Z, Skogerbø G, chen L, Lu H, Zhao Y, Chen R: NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Res 2006, 34:D150-2.

CHAPTER 2. COMPUTATIONAL TOOLS FOR INVESTIGATING RNA-PROTEIN INTERACTION PARTNERS

Modified from a paper published in Journal of Computer Science and Systems Biology, 2013, 6:182-187

Usha K Muppirala, Benjamin A Lewis and Drena Dobbs

Abstract

RNA-protein interactions are important in a wide variety of cellular and developmental processes. Recently, high-throughput experiments have begun to provide valuable information about RNA partners and binding sites for many RNA-binding proteins (RBPs), but these experiments are expensive and time consuming. Thus, computational methods for predicting RNA-protein interactions (RPIs) can be valuable tools for identifying potential interaction partners of a given protein or RNA, and for identifying likely interfacial residues in RNA-protein complexes. This review focuses on the “partner prediction” problem and summarizes available computational methods, web servers, and databases that are devoted to it. New computational tools for addressing the related “interface prediction” problem are also discussed. Together, these computational methods for investigating RNA-protein interactions provide the basis for new strategies for integrating RNA-protein interactions into existing genetic and developmental regulatory networks, an important goal of future research.

Introduction

In the post-transcriptional regulation of gene expression, RNA-binding proteins (RBPs) interact with target mRNAs and non-coding RNAs (ncRNAs) to regulate a variety of cellular processes including RNA splicing, RNA transport and stability, and translation (Kishore *et al.,* 2010, Licatalosi *et al.,* 2010, Singh *et al.,* 20012). RNA-protein interactions (RPIs) also play important roles in human health and diseases (Khalil *et al.,* 2011) as well as in viral replication (Li *et al.,* 2011) and pathogen resistance in plants (Zvereva *et al.,* 2012). Even though the human genome contains more than 400 known or predicted RBPs (Ray *et al.,* 2013, cook *et al.,* 2011), the structures of RNA-protein complexes and the roles of RPIs in post-transcriptional regulatory networks (Kishore *et al.,* 2010, Mittal *et al.,* 2009) are much less well characterized than the DNA-protein complexes involved in transcriptional regulation. For example, on July 18, 2013, the Protein Data Bank (PDB) (Berman *et al.,* 2000) contained only 1,593 structures of RNA-protein complexes, compared with more than 2,800 structures of DNA-protein complexes. Recently, however, new experimental approaches have been used to interrogate RNA-protein complexes and interaction networks. For example, high-throughput in vivo and in vitro experiments have been used to identify cellular RNA molecules that bind a protein of interest (Ankö *et al.,* 2012, König *et al.,* Riley *et al.,* 2013). Global proteomic approaches have been applied to identify the entire mRNA-bound proteome (Baltz *et al.,* 2012).

The available structures of RNA-protein complexes in the PDB, databases of protein and RNA motifs, and a growing knowledge base regarding RNA and protein interactions in the literature have been exploited to develop computational methods for addressing several questions about RNA-protein interactions:

- Does this protein bind RNA?
- Which RNA molecules are bound by this protein?
- Which RNA sequence or structural motifs are recognized by this protein?
- Which amino acid residues are directly involved in binding RNA?

In this review, we focus on existing computational methods and web servers for predicting RNA-protein interaction partners. We also discuss recently developed "partner-aware" approaches for predicting RNA-protein interfaces, which use information about both the protein and RNA molecules to identify binding regions in either one or both sequences. Finally, available curated databases of RNA-protein interactions are briefly reviewed.

RNA-Protein Partner Prediction Methods and Web Servers

Table 2.1 summarizes the characteristics of computational methods available for predicting the interaction probability of a given RNA-protein pair. A general description of the machine learning methods and performance metrics discussed below is provided in Supplementary Text S1.

Table 2.1 Computational methods for predicting RNA-protein interaction partners

Method	Dataset	Features	Description
Pancaldi and Bähler	5,166 mRNA-protein interacting pairs from immunopurification experiments	Predicted protein secondary structure, localization, protein physical properties, gene physical properties, UTR properties, genetic interactions	Protein and RNA sequences encoded using > 100 features are used to train SVM and RF classifiers
Bellucci et al. (catRAPID)	410 interacting pairs from 858 RNA-protein complexes from PDB	Physicochemical properties including secondary structure propensities, hydrogen-bonding propensities, and van der Waals interaction propensities	Propensities are calculated for each amino acid and ribonucleotide to generate an interaction profile (http://service.tartaglialab.com/page/catrapid_group)

Table 2.1 (continued)

Method	Dataset	Features	Description
Muppirala et al. (RPISeq)	2,241 interacting pairs from 943 RNA-protein complexes from PRIDB (RPI2241)	Sequence composition of proteins, represented as conjoint triads, and RNAs, represented as tetrads	Protein and RNA sequences encoded sequence-composition-based features are used to train SVM and RF classifiers (http://pridb.gdcb.iastate.edu/RPISeq)
Wang et al.	RPI 2241 generated by Muppirala et al. & 367 interacting pairs from NPInter	Sequence composition of protein and RNA	Input to NB and ENB classifiers is a combination of protein triads and RNA triad features similar to those used in RPISeq

To the best of our knowledge, the first method for computationally predicting mRNA-protein interactions was proposed by Pancaldi and Bähler in 2011 (Pancaldi and Bähler, 2011) . Their study took advantage of a dataset of 5,166 mRNA-RBP interactions detected using RNA immunopurification experiments performed in S. cerevisiae (Hogan *et al.,* 2008). Two machine learning methods, Support Vector Machines (SVMs) and Random Forest (RF) classifiers (see Supplementary Text S1), were used to predict the likelihood of interaction between an RBP and its target mRNAs. Input for the classifiers included more than 100 characteristic gene and protein features, but no motifs or experimentally measured binding specificities were used. Feature classes included gene ontology terms, predicted secondary structures, mRNA properties, and genetic interactions. Overall, the RF classifier performed slightly better than SVM. In 2-fold cross validation experiments, an average prediction accuracy of 69% was obtained, with average sensitivity of 70% and specificity of 69%. When the authors tried to predict the mRNA targets of individual RBPs that were not included in the training set, the performance of the classifiers was highly variable across the RBPs. On average, the classifiers performed with an accuracy of only 50%. Using the pre-rRNA processing factor Nop15p as an example, the authors demonstrated that their method performs better when the training set includes at least some of the known mRNA targets for a

given RBP. The authors acknowledge that the main limitation of this method is that it requires many features of both the RNA and protein under consideration. Although some of these features are easy to compute, some of them may not be available for other RNA-protein pairs of interest, and they are not trivial to obtain experimentally. Hence, the method may have limited applicability.

Also in 2011, the catRAPID method for predicting long non-coding RNA (lncRNA) partners of RBPs was published (Bellucci *et al.,* 2011). This study used a dataset consisting of 858 RNA-protein complexes extracted from the PDB (Berman *et al.,* 2000). Values for several physicochemical properties, including secondary structure propensities, hydrogen bonding propensities and van der Waals interaction propensities, were combined to calculate an interaction profile for each lncRNA and protein, which was then used to calculate interaction propensities for every potential lncRNA-protein pair. The interaction propensity of a RNA-protein pair in the training dataset was reported using the discriminative power (DP), which ranges between 0 and 1, with higher confidence interactions having higher DP values. The reported discriminative power on a non-redundant training set was 78%. The performance of catRAPID was also evaluated on independent test sets composed of positive interactions from the NPInter database of ncRNA-protein interactions (Wu *et al.,* 2006), for which 89% prediction accuracy was reported (Bellucci *et al.,* 2011). However, when tested on 12,000 randomly generated RNA associations with proteins extracted from a non-Nucleic Acid-binding dataset (Stawiski *et al.,* 2003), ~30% of these were predicted to interact with RNA (bellucci *et al.,* 2011). In a recent study (cirillo *et al.,* 2013), the authors used catRAPID to investigate ribonucleoprotein interactions linked to neurodegenerative diseases. An advantage of the catRAPID algorithm is that it is the only published method that

simultaneously predicts the binding sites in both RNA and protein sequences (Cirillo *et al.*, 2013). The catRAPID web server is available at http://service.tartaglialab.com/page/catrapid_group.

A purely sequence-based approach to predict RPIs, RPISeq, was proposed by our group, also in 2011 (Muppirala *et al.*, 2011). RPISeq is a family of machine learning classifiers (RF and SVM) designed to predict the probability of interaction between a given protein and RNA. In this method, RNA sequences are encoded as normalized frequencies of RNA tetrads, and protein sequences are encoded using a conjoint triad feature (CTF) method originally proposed by Shen et al. (2007). In essence, RPISeq exploits the amino acid composition of protein sequences and ribonucleotide composition of RNA sequences to predict the probability that a given pair (one protein and one RNA) will interact. On a non-redundant dataset of 2241 interacting pairs (RPI2241) created from known RNA-protein complexes in PRIDB (Lewis *et al.*, 2011), the RPISeq-RF classifier performed slightly better (average accuracy 89.6%), compared to the RPISeq-SVM classifier (average accuracy 87.1%). On an independent test set composed of only positive examples generated from NPInter, the RPISeq-RF classifier correctly predicted 80.2% of interactions, while RPISeq-SVM predicted 66.3% of interactions. RPISeq's performance on an independent negative dataset was not reported. RPISeq's performance, using sequence information alone, was comparable to that of Pancaldi and Bähler's method, which uses extensive feature information. An independent experimental validation of RPISeq predictions was published in a recent study (He *et al.*, 2013), in which RPISeq was used to predict that the linc-UBC1 RNA interacts with PRC2 (Polycomb Repressive Complex 2). This prediction was experimentally validated using RNA immunoprecipitation, which confirmed that linc-UBC1

physically interacts with two core protein components of the PRC2 complex, EZH2 and SUZ12. RPISeq is available as a web server at http://pridb.gdcb.iastate.edu/RPISeq.

Another sequence-based method, similar to RPISeq, was proposed by Wang et al. in 2012 (Wang *et al.*, 2013). This study also used the RPI2241 dataset (Muppirala *et al.*, 2011) as one of the training datasets, a variation of the conjoint triad feature representation as protein descriptors and frequencies of RNA triads as RNA descriptors. The feature vector also included all combinations of protein and RNA descriptors. Only those features that were enriched in the training dataset were used as input for Naïve Bayes (NB) and Extended Naïve Bayes (ENB) classifiers (see Supplementary Text S1). In cross-validation experiments using the RPI2241 dataset, the ENB classifier had a slightly better accuracy than the NB classifier (74% vs. 73%). The classifiers were also evaluated on known interactions from an independent dataset extracted from NPInter, with a reported predictive power of 79% (using the ENB classifier trained on RPI2241). In another experiment, the authors used a dataset of 30 ncRNAs and 759 proteins to predict RNA-protein interactions in C. elegans. They used an ncRNA pull-down experiment to validate these predictions for one selected ncRNA, sbRNA CeN72. The experiments identified 51 proteins that interact with CeN72. However, the ENB classifier predicted a total of 207 CeN72 interacting proteins (see Supplemental Table S5 in (Wang *et al.*, 2013)); of these, only 10 were true positive predictions. Although the authors claim that their method outperforms other existing methods, no evidence was presented to support this claim. In fact, as summarized in Supplementary Table 2.S1, the published results demonstrate that RPISeq-RF (Muppirala *et al.*, 2011) outperforms the ENB classifier (Wang *et al.*, 2013).

In summary, except for Pancaldi and Bähler's approach, all of the methods discussed above use sequence information as the primary input to make predictions. This is a distinct advantage when making predictions on proteins or RNAs for which little information is available, other than the sequence. Also, every method except that of Pancaldi and Bähler uses training data partly derived from three-dimensional structures of complexes in the PDB. Because the number of experimentally determined structures of RNA-protein complexes is relatively small and the PDB does not yet encompass all possible types of RNA-protein interactions, one should use caution when interpreting these predictions. A weakness of all of these predictors is the use of a negative dataset generated from random pairings of RNAs and proteins (in which many false negative examples may be included). Using real negative examples based on experimental interaction data would be desirable and would increase confidence in the predictions.

In conclusion, researchers interested in predicting RPIs are advised to compare results of more than one method. At present, only two of the methods described above are available as web-based servers (see Table 2.1).

Web servers for partner prediction

The catRAPID server (http://service.tartaglialab.com/page/catrapid_group) developed by Bellucci et al. (2011) provides an estimate of the interaction propensities of given RNA and protein sequences. The output is displayed as a heat-map of interaction scores, with x and y axes representing the RNA and protein sequences, respectively. The overall interaction score and the corresponding discriminative power (predictive measure for binding) are also reported. This server provides another module called catRAPID strength that predicts the

"strength" of a RNA-protein pair by comparing its interaction propensity with the interaction propensities of a reference set of 100 proteins and 100 RNAs.

The RPISeq web server (http://pridb.gdcb.iastate.edu/RPISeq) implements the RPISeq method developed by Muppirala et al. (2011). RPISeq takes as input a pair of RNA and protein sequences and outputs the interaction probability computed by SVM and RF classifiers trained using the RPI2241 dataset. It also accepts batch submission of multiple proteins or RNAs. Currently, users can input a maximum of 100 sequences. This limitation can be overcome by using a stand-alone version of the program, which is freely available from the authors.

RNA-Protein Interface Prediction Methods

So far, we have discussed computational methods for predicting the likelihood that a given RNA-protein pair will interact. Understanding how individual RNAs and proteins specifically recognize each other is an important aspect of this problem, and requires characterization of interfacial contacts at the residue and atomic level. As a step toward deciphering the rules that govern recognition specificity in RNA-protein interfaces, many computational methods (both sequence-based and structure-based) have been developed for predicting RNA-binding residues in proteins. Three recent reviews have summarized and compared these methods (Cirillo *et al.*, 2013, Puton *et al.*, 2012, Walia *et al.*, 2012), which we will not reconsider here. With one exception, all published methods for predicting RNA-binding residues in a protein of interest do not take into account the specific RNA partner with which it interacts (i.e., they are or "partner-agnostic" or "non-partner specific" methods. Here, we will focus instead on methods that are "partner-aware" or "partner-specific." For

protein-protein complexes, the partner-specific approach has been shown to provide improved interface predictions over non-partner specific methods in several studies (e.g., (Ahmad and Mizuguchi, 2011, Xue *et al.*, 2011)).

The first partner-specific RNA-binding residue prediction method was proposed by the Han group (Shrestha *et al.*, 2008, Choi and Han, 2011). In this work, both protein and RNA features were used as input to an SVM classifier to predict RNA-binding residues. Length and amino acid composition of the protein, along with features such as solvent accessible surface area and interaction propensity of an amino acid triplet were used to encode the input protein. The input RNA was encoded as a 4 element vector representing the sum of the normalized position of each ribonucleotide in the RNA sequence. In 5-fold cross-validation experiments on a dataset of 3,149 RNA-protein interacting pairs, prediction accuracy was 84%, with a correlation coefficient (CC) 0.41. On an independent dataset comprising 267 RPIs, accuracy was 90%, with CC of 0.24 (Agostini *et al.*, 2013). Comparison with non-partner specific methods on the same datasets showed that the performance of the partner-specific approach was superior in terms of CC, and comparable in terms of overall accuracy. It seems likely that using more descriptive features to encode the sequence of the RNA partner could provide improved performance.

A second partner-specific prediction method for identifying binding sites in both the protein and RNA partners of an interacting pair is catRAPID (Bellucci *et al.*, 2011). As discussed above, catRAPID predicts interaction partners based on the interaction propensities of individual residues (Bellucci *et al.*, 2011). In several cases, catRAPID binding site predictions correlate well with experimental results (Cirillo *et al.*, 2013, Agostini *et al.*, 2013), but the performance of this method has not been evaluated systematically on

benchmark datasets. Therefore, it is difficult to comment on the relative accuracy of this method in predicting interfacial residues in either RNA or protein sequences.

Sequence and structural motifs in RNA-protein interfaces

Structural analyses of RNA-protein complexes and sequence data from high-throughput RNA-protein interaction experiments have led to a rapid expansion in the collections of structural and sequence motifs associated with interfaces in RNA-protein complexes. Databases of protein motifs (e.g., ProSite (Sigrist *et al.,* 2010)) and RNA motifs (e.g., FR3D (Sarver *et al.,* 2008)) are valuable resources for investigating recognition principles in RNA-protein interactions. In addition to their utility for identifying binding sites in novel proteins and RNAs, motifs can provide insight into the biological functions of protein or RNA families.

Well-characterized RNA-binding motifs in proteins include the RNA recognition motif (RRM), the K-homology (KH) domain, the Pumilio/FBF (PUF) domain, and the double-stranded RNA-binding domain (dsRBD) (recently reviewed in Chen and Varani, 2013). The number of characterized RNA structural motifs is smaller, but includes several well-studied examples, such as pseudoknots, tetra-loops, and kink turns (Fritsch and Westhof, 2010). RNA sequence motifs that serve as recognition sites for RBPs have been identified using in vitro selection methods such as SELEX (Tuerk and Gold, 1990) and RNAcompete (Ray *et al.,* 2009). High-throughput approaches for capturing in vivo RNA-protein complexes by Tap-tagging and immunoprecipitation (Hogan *et al.,* 2008) or UV crosslinking and immunoprecipitation of RNA-protein complexes combined with microarray or RNA-Seq analysis (Ankö *et al.,* 2012, König *et al.,* 2011) have resulted in a dramatic

increase in our understanding of recognition motifs in cellular RNAs. Experimental data from such studies have been analyzed to determine sequence and structural features of recognition motifs for RBPs using methods such as RNAcontext (Kazan *et al.*, 2010). These data are now available in resources such as the RBPDB database (Cook *et al.*, 2011) (see below), and in RBPMotif (Kazan *et al.*, 2013), a web server for identifying sequence and structure preferences of RBPs.

RNA-Protein Interaction Databases

At present, there is no single comprehensive database of RNA-protein interactions. Widely used databases that contain RNA-protein complexes and/or interactions as part of a broader collection include structure databases, such as the PDB (Berman *et al.*, 2000) and NDB (Berman *et al.*, 1992), as well as interaction databases, such as BioGRID (Stark *et al.*, 2011) and IntAct (Kerrien *et al.*, 2012). The Protein Data Bank (PDB) is a comprehensive database of experimentally determined three-dimensional structures of macromolecules, including both proteins and nucleic acids. The Nucleic Acid Database (NDB) contains experimental 3D structural information for nucleic acids, and includes both DNA-protein and RNA-protein complexes. BioGRID is a curated database of protein interactions and genetic interactions from more than 45 model organisms. The IntAct database primarily contains protein-protein interactions, although it also includes some protein-small molecule, protein-nucleic acid and protein-gene locus interactions.

In the remainder of this section, several databases that focus on RNA-protein interactions are discussed. Table 2.2 provides URLs for these.

The first three databases discussed below, PRD (Fujimori *et al.*, 2012), NPInter (Wu *et al.*, 2006) and RPIntDB (http://pridb.gdcb.iastate.edu/RPISeq/) are collections of RNA-protein interaction partners. They focus on binary interactions between proteins and RNAs and do not provide residue or atomic level information about interfaces. Most interactions in these databases are extracted from results of low-throughput, or more recently, high-throughput experiments in published literature.

In contrast, PRIDB (http://pridb.gdcb.iastate.edu) (Lewis *et al.*, 2011) is a collection of interfaces in RNA-protein complexes, derived from experimentally determined structures deposited in the PDB. Databases similar to PRIDB, but not focused exclusively on RNA-protein complexes, include ProNIT (http://www.abren.net/pronit/) (Kumar *et al.*, 2006), which contains experimentally determined thermodynamic interaction data for protein-nucleic acid interactions; BIPA (http://mordred.bioc.cam.ac.uk/bipa) (Lee *et al.*, 2009), the Biological Interaction Database for Protein-Nucleic Acid; and NPIDB (http://npidb.belozersky.msu.ru) (Kirsanov *et al.*, 2013), which also includes structural information for both DNA-protein and RNA-protein complexes, as well as several online tools for analysis.

The final database included in this section, RDPDB (Ray *et al.*, 2013, Cook *et al.*, 2011), is a recently expanded collection of RNA-binding proteins and their experimentally determined target RNAs. This database provides information about both RNA-protein interaction partners and their interfaces, with a focus on the RNA recognition preferences of individual RPBs.

Table 2.2 Databases of RNA-protein interactions and interfaces.

Database	URL	Description
BioGRID	http://thebiogrid.org/	Manually curated protein and genetic interactions for major model organisms
IntAct	http://www.ebi.ac.uk/intact/	Manually curated molecular interactions, including comprehensive data about their source experiments
NDB	http://ndbserver.rutgers.edu/	Nucleic acid and DNA/RNA-protein complex structures, including derived data for nucleic acids
NPInter	http://www.panrna.org/NPInter/index.php	Functional interactions of ncRNAs and protein-related biomolecules, classified into categories based on interaction type
PDB	http://www.rcsb.org/pdb/home/home.do	Experimentally determined three-dimensional structures
PRD	http://pri.hgc.jp/	RPIs from 22 species, focusing on gene-level information
PRIDB	http://pridb.gdcb.iastate.edu/	Interface information from RNA-protein complex structures in browsable and machine-readable format
RBPDB	http://rbpdb.ccbr.utoronto.ca/	Experimental data on binding preferences and specificities of RBPs
RPIntDB	http://pridb.gdcb.iastate.edu/RPISeq/	RPIs from databases and high-throughput experiments in literature

PRD

PRD (http://pri.hgc.jp/) (Fujimori *et al.*, 2012) is the most comprehensive database of RNA-protein interactions currently available. It contains more than 10,000 documented physical interactions between RNA and proteins. It includes interactions from BioGRID, IntAct and the PDBj (Kinjo *et al.*, 2012). The PRD interaction data model is based on the HUPO POSI-MI model and the database can be searched using 11 different fields (e.g., Gene ID, experiment, biological function) or using text keywords. Each interaction record contains information about both the protein and RNA involved, the experimental method used to detect the interaction, and references. Biological functions and information regarding binding sites are also provided, when available. Search results can be exported in PSI-MI XML files.

NPInter

NPInter (http://www.panrna.org/NPInter/index.php) (Wu *et al.*, 2006) was the first database developed to collect experimentally determined functional interactions between ncRNAs and protein-related biomolecules (PRMs), i.e., proteins, mRNAs or genomic DNAs. Interactions involving tRNAs and rRNAs are not included. In 2006, NPInter contained 700 interactions from six model organisms. NPInter version 2.0, available in 2013, now contains more than 200,000 interactions from 18 different organisms. It classifies the interactions into eight categories: 'ncRNA binds protein', 'ncRNA regulates mRNA expression', 'ncRNA indirectly regulates a gene activity', 'ncRNA expression is regulated by protein', 'ncRNA affects protein activity', 'ncRNA activity is affected by protein', 'genetic interaction between ncRNA gene and protein gene' and 'other linkages'. Users can search NPInter by molecule type (ncRNA, miRNA, protein) by ID (NONCODE, miRBase, UniProt, PubMed), or using text queries. NPInter provides a BLAST option to query protein, ncRNA, and miRNA sequences. Multiple download options are also provided.

RPIntDB

The RNA-Protein Interaction DataBase (RPIntDB), (http://pridb.gdcb.iastate.edu/RPISeq/) was developed as a component of the RPISeq server (Muppirala *et al.*, 2011). The database includes experimentally validated RNA-protein interactions from several sources. It includes 11,815 proteins and 2,408 RNAs extracted from known RNA-protein complexes in PRIDB (as of March 2011), 242 ncRNAs and 282 proteins from ncRNA-protein interactions in the NPInter database and 13,243 RPIs from high-throughput experiments published in literature (Hogan *et al.*, 2008). Users can query

RPIntDB to determine whether there is experimental evidence that a specific protein of interest is involved in an RPI. In the current version of RPIntDB, the service runs a BLAST search against the database and returns protein sequences that fall within a user-specified e-value threshold, along with their experimentally validated interacting RNA partners. The corresponding source(s) of the interaction are displayed in the output results.

PRIDB

The Protein-RNA Interface Database (PRIDB) (http://pridb.gdcb.iastate.edu) (Lewis *et al.*, 2011) is a comprehensive database of RNA-protein interfaces extracted from RNA-protein complexes in the PDB. It contains 16,350 proteins and 3,398 RNAs from 1,484 RNA-protein complexes (as of July 1 2013). PRIDB displays interfacial residues on protein and RNA sequences. It also displays known RNA-binding domains or motifs from ProSite (Sigrist *et al.*, 2010) and RNA structural motifs from FR3D (Sarver *et al.*, 2008). Atomic-level contact details for interfaces in the RNA-protein complexes can be visualized using an integrated JMol applet or downloaded in a machine-readable format. PRIDB also provides several reduced-redundancy benchmark datasets of RNA-binding protein chains.

RBPDB

The RNA-Binding Protein Database (RBPDB) (http://rbpdb.ccbr.utoronto.ca) (Ray *et al.*, 2013, Cook *et al.*, 2011) is a highly valuable compendium of experimentally determined RNA-binding specificities for RBPs from human, mouse, D. melanogaster and C. elegans. RBPDB contains target site preferences for more than 200 RBPs, extracted from almost 1,500 RNA-binding experiments. RBPDB catalogues data from 14 types of RNA-binding

experiments and includes binding site sequence logos for more than 70 RBPs. The database can be searched by RBD, experiment type, species and gene name.

Future Directions

The emergence of high-throughput experimental approaches for interrogating RNA-protein interactions is generating a vast amount of new data, which will undoubtedly lead to improved computational methods for analyzing and predicting RNA-protein interfaces and interaction partners.

Despite recent advances in both experimental and computational methodology, identifying the interaction partner(s) for a specific protein or RNA sequence is still an immensely challenging task. For example, even though the compendium of RNA-binding proteins and their targets published by the Hughes and Morris laboratories includes RBP recognition sites for more than 200 different RBPs (Ray *et al.*, 2013), this impressive number corresponds to less than half of the known RBPs encoded in the human genome (cook *et al.*, 2011). An analysis of the mRNA-bound proteome of a human kidney cell line identified ~800 bound proteins (Baltz *et al.*, 2012), nearly one third of which were not previously annotated as RNA-binding. With such large numbers of RPBs, each of which binds multiple mRNA and/or ncRNA targets, another difficult task will be to identify which combinations of RBPs determine specific post-transcriptional fates of individual mRNAs and ncRNAs. Progress in this direction was demonstrated in a quantitative proteomic analysis in S. cerevisiae, which identified sets of RBPs that bind simultaneously to common RNA targets (Klass *et al.*, 2013). Computational tools for constructing and interrogating RNA-protein

interaction networks and for integrating RPIs into existing gene and protein interaction networks will be needed.

Obtaining high-resolution experimental structures of RNA-protein complexes is notoriously difficult and time consuming (Ke and Doudna, 2004, Scott and Hennig, 2008). Thus, improved methods for computational modeling will be important for gaining insight into molecular details of interfaces in recalcitrant RNA-protein complexes. Algorithms for RNA-protein docking (not discussed in this review), although still somewhat naïve relative to those for small molecule and protein docking, are already benefitting from the increased availability of RNA-containing complex structures (Tuszynska and Bujnicki, 2011, Li *et al.*, 2012, Huang *et al.*, 2013).

Finally, another important future direction in research on RNA-protein interactions is the rational design of RNA-protein interfaces. Engineered DNA binding proteins, such as ZFNs and TALENS, have become enormously powerful tools for genome engineering and are poised to enter clinical settings (Joung and Sander, 2013, Reyon *et al.*, 2012, Rahman *et al.*, 2011). Likewise, RNA-binding proteins engineered to recognize specific RNA sequences (Chen and Varani, 2013) could become valuable tools for manipulating post-transcriptional regulatory networks in the research laboratory, and potentially, important therapeutic agents for treating genetic and infectious diseases.

Acknowledgements and Funding

We thank Rasna Walia, Xue Li and Pete Zaback for suggestions and critical comments on the manuscript. This work was partially supported by funding from National Institutes of Health (GM066387 to D.D.).

Supplementary materials

Supplementary Text S1: Machine learning methods and evaluation metrics discussed in this review.

Machine learning offers one of the most cost-effective approaches to constructing predictive models in settings where experimentally validated training data are available (Mitchell T, 1997, Machine Learning). This review focuses on machine learning classifiers that are designed to predict whether or not a given input RNA-protein pair interacts. The classifiers are trained using experimentally validated pairs of RNAs and proteins, together with their interaction "classification," i.e., interacting (positive) or non-interacting (negative). When negative examples are not available, it is common practice to randomly generate negative training data. Trained classifiers are used to make predictions on unknown RNA-protein pairs. For every instance or subject, the classifier outputs a probability value ranging from 0 to 1. Instances above a certain threshold (typically 0.5) are classified as "interacting," and instances below the threshold are classified as "non-interacting". Different researchers have used: i) different datasets for training; ii) different types of information about the RNAs and proteins as input, e.g., primary sequences or structural features; iii) different encodings of the input information; and v) different types of machine learning classifiers. A brief description of the four types of machine learning classifiers discussed in this review follows.

Machine Learning Classifiers

For an in depth treatment, see Mitchell (Mitchell T, 1997, Machine Learning).

Naïve Bayes (NB) classifier (Mitchell T, 1997, Machine Learning) is a probabilistic classifier based on Bayesian statistics. It makes the simplifying assumption that all attributes

are independent given the class label. The **ENB classifier** used in Wang et al. (Wang et al., 2013, Molecular BioSystems 9:133-42) is an extension of the standard NB model, in which the correlation between features is considered.

Random Forest (RF) (Breiman I, 2001, Machine Learning 45:5-32) is an ensemble of classification trees. Each classification tree in the ensemble is trained on a subset of training examples that are randomly sampled from the entire training set. At each node, the best split is chosen from a set of *m* variables selected at random from the set of input features. Given a query instance, the majority vote of all the classifiers is returned as the RF prediction.

Support Vector Machine (SVM) (Vapnik V, 1995, The Nature of Statistical Learning Theory) classifies data by finding a hyperplane that maximizes the margin of separation between two classes. A strength of SVMs is that they can distinguish classes that are not linearly separable by mapping the input onto a higher-dimensional space using a kernel function.

Performance Evaluation Metrics

The performance of a classifier can be summarized in four output measures.

TP (true positives) = the number of positive examples correctly predicted as positives

TN (true negatives) = the number of negative examples correctly predicted as negatives

FP (false positives) = the number of negative examples incorrectly predicted as positives

FN (false negatives) = the number of positive examples incorrectly predicted as negatives

Based on these values, commonly calculated performance metrics include sensitivity, specificity, precision, accuracy, F-measure and correlation coefficient. These terms are described and defined below.

Sensitivity or ***recall*** is a measure of the classifier's ability to identify positive examples. It is defined as the fraction of actual positives that are predicted to be positives.

Specificity relates to the classifier's ability to identify negative examples. It is defined as the fraction of actual negatives that are predicted to be negatives. In the papers discussed in the review, specificity was calculated as per the above definition. According to Baldi (Baldi P et al., 2000, Bioinformatics 16:412-424), Specificity is alternatively defined as the probability that a positive prediction is correct. This alternate definition is also widely used in classifier assessments.

Precision is the fraction of predicted positives that are true positives.

Accuracy is the percentage of the correctly predicted positive and negative examples.

F-Measure is the harmonic mean of precision and recall.

Matthews Correlation Coefficient (MCC) measures the linear correlation between the actual and predicted binary classification.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F\text{-}Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Comparison of RPISeq classifier with Wang et al.'s method

Supplementary Table 2.S1: Published performance metrics for the ENB classifier of Wang et al. (2013) *Molecular BioSystems* 9:133 and the RPISeq-RF classifier of Muppirala et al. (2011) *BMC Bioinformatics* 12:489, on a balanced RPI2241 dataset.

Performance Metrics	ENB Classifier	RPISeq-RF Classifier
Accuracy	0.67	0.90
Sensitivity	0.56	0.90
Specificity	0.79	0.89*
Precision	0.73	0.89
MCC	0.36	0.79*

Both the classifiers used a balanced dataset including 2,241 positive examples and 2,241 negative examples extracted from the RPI2241 dataset. The specific negative examples used in the two studies may differ. The results reported were obtained using 10-fold cross validation experiments. *The specificity and MCC values for RPISeq were not included in the cited publication.

References

Agostini F, Cirillo D, Bolognesi B, Tartaglia GG: X-inactivation: quantitative predictions of protein interactions in the Xist network. Nucleic Acids Research 2013, 41:e31.

Ahmad S, Mizuguchi K: Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. PLoS ONE 2011, 6:e29104.

Ankö M-L, Neugebauer KM: RNA-protein interactions in vivo: global gets specific. Trends in Biochemical Sciences 2012, 37:255–62.

Baltz AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, et al: The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular Cell 2012, 46:674–90.

Bellucci M, Agostini F, Masin M, Tartaglia GG: Predicting protein associations with long noncoding RNAs. Nature Methods 2011, 8:444–5.

Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, et al: The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophysical Journal 1992, 63:751–759.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al: The Protein Data Bank. Nucleic Acids Research 2000, 28:235–42.

Chen Y, Varani G: Engineering RNA-binding proteins for biology. The FEBS Journal 2013, doi: 10.1111/febs.12375.

Choi S, Han K: Prediction of RNA-binding amino acids from protein and RNA sequences. BMC Bioinformatics 2011 12 Suppl 1:S7.

Cirillo D, Agostini F, Klus P, Marchese D, Rodriguez S, et al: Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. RNA 2013, 19:129–40.

Cirillo D, Agostini F, Tartaglia GG: Predictions of protein-RNA interactions. Wiley Interdisciplinary Reviews: Computational Molecular Science 2013, 3:161–175.

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR: RBPDB: a database of RNA-binding specificities. Nucleic Acids Research 2011, 39:D301–8.

Fritsch V, Westhof E: The Architectural Motifs of Folded RNAs, in The Chemical Biology of Nucleic Acids. John Wiley & Sons, Ltd, Chichester, UK 2010.

Fujimori S, Hino K, Saito A, Miyano S, Miyamoto-Sato E: PRD: A protein-RNA interaction database. Bioinformation 2012, 8:729–30.

He W, Cai Q, Sun F, Zhong G, Wang P, et al: linc-UBC1 physically associates with polycomb repressive complex 2 (PRC2) and acts as a negative prognostic factor for lymph node metastasis and survival in bladder cancer. Biochimica et Biophysica Acta 2013, 1832:1528–1537.

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO: Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biology 2008, 6:e255.

Huang Y, Liu S, Guo D, Li L, Xiao Y: A novel protocol for three-dimensional structure prediction of RNA-protein complexes. Scientific Reports 2013, 3:1887.

Joung JK, Sander JD: TALENs: a widely applicable technology for targeted genome editing. Nature Reviews Molecular Cell Biology 2013, 14:49–55.

Kazan H, Morris Q: RBPmotif: a web server for the discovery of sequence and structure preferences of RNA-binding proteins. Nucleic Acids Research 2013, 41:W180–6.

Kazan H, Ray D, Chan ET, Hughes TR, Morris Q: RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Computational Biology 2010, 6:e1000832.

Ke A, Doudna JA: Crystallization of RNA and RNA-protein complexes. Methods 2004, 34:408–14.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, et al: The IntAct molecular interaction database in 2012. Nucleic Acids Research 2012, 40:D841–6.

Khalil AM, Rinn JL: RNA-protein interactions in human health and disease. Seminars in Cell and Developmental Biology 2011, 22:359–65.

Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, et al: Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Research 2012, 40:D453–60.

Kirsanov DD, Zanegina ON, Aksianov EA, Spirin SA, Karyagina AS, et al: NPIDB: Nucleic acid-Protein Interaction DataBase. Nucleic Acids Research 2013, 41:D517–23.

Kishore S, Luber S, Zavolan M: Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. Briefings in Functional Genomics 2010, 9:391–404.

Klass DM, Scheibe M, Butter F, Hogan GJ, Mann M, et al: Quantitative proteomic analysis reveals concurrent RNA-protein interactions and identifies new RNA-binding proteins in Saccharomyces cerevisiae. Genome Research 2013, 23:1028–38.

König J, Zarnack K, Luscombe NM, Ule J: Protein-RNA interactions: new genomic technologies and perspectives. Nature Reviews Genetics 2011, 13:77–83.

Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, et al: ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Research 2006, 34:D204–6.

Lee S, Blundell TL: BIPA: a database for protein-nucleic acid interaction in 3D structures. Bioinformatics 2009, 25:1559–60.

Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, et al: PRIDB: a Protein-RNA interface database. Nucleic Acids Research 2011, 39:D277–82.

Li CH, Cao L Bin, Su JG, Yang YX, Wang CX: A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. Proteins 2012, 80:14–24.

Li Z, Nagy PD: Diverse roles of host RNA binding proteins in RNA virus replication. RNA Biology 2011, 8:305–15.

Licatalosi DD, Darnell RB: RNA processing and its regulation: global insights into biological networks. Nature Reviews Genetics 2010, 11:75–87.

Mittal N, Roy N, Babu MM, Janga SC: Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. Proceedings of the National Academy of Sciences of the United States of America 2009, 106:20300–5.

Muppirala UK, Honavar VG, Dobbs D: Predicting RNA-protein interactions using only sequence information. BMC Bioinformatics 2011, 12:489.

Pancaldi V, Bähler J: *In silico* characterization and prediction of global protein-mRNA interactions in yeast. Nucleic Acids Research 2011, 39:5826–36.

Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM: Computational methods for prediction of protein-RNA interactions. Journal of Structural Biology 2012, 179:261–8.

Rahman SH, Maeder ML, Joung JK, Cathomen T: Zinc-finger nucleases for somatic gene therapy: the next frontier. Human Gene Therapy 2011, 22:925–33.

Ray D, Kazan H, Chan ET, Peña Castillo L, Chaudhry S, et al: Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nature Biotechnology 2009, 27:667–70.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, et al: A compendium of RNA-binding motifs for decoding gene regulation. Nature 2013, 499:172–177.

Reyon D, Tsai SQ, Khayter C, Foden JA, Sander JD, et al: FLASH assembly of TALENs for high-throughput genome editing. Nature Biotechnology 2012, 30:460–5.

Riley KJ, Steitz JA: The “Observer Effect” in genome-wide surveys of protein-RNA interactions. Molecular Cell 2013, 49:601–4.

Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB: FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. Journal of Mathematical Biology 2008, 56:215–52.

Scott LG, Hennig M: RNA structure determination by NMR. Methods in Molecular Biology 2008, 452:29–61.

Shen J, Zhang J, Luo X, Zhu W, Yu K, et al: Predicting protein-protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America 2007, 104:4337–41.

Shrestha R, Kim J, Han K: Prediction of RNA-Binding Residues in Proteins Using the Interaction Propensities of Amino Acids and Nucleotides, in Advanced Intelligent Computing Theories and Applications. Springer-Verlag Berlin Heidelberg 2008.

Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, et al: PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Research 2010, 38:D161–6.

Singh R: RNA-protein interactions that regulate pre-mRNA splicing. Gene Expression 2002, 10:79–92.

Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R, et al: The BioGRID Interaction Database: 2011 update. Nucleic Acids Research 2011, 39:D698–704.

Stawiski EW, Gregoret LM, Mandel-Gutfreund Y: Annotating nucleic acid-binding function based on protein structure. Journal of Molecular Biology 2003, 326:1065–79.

Tuerk C, Gold L: Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 1990, 249:505–10.

Tuszynska I, Bujnicki JM: DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. BMC Bioinformatics 2011, 12:348.

Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, et al: Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinformatics 2012, 13:89.

Wang Y, Chen X, Liu Z-P, Huang Q, Wang Y, et al: *De novo* prediction of RNA-protein interactions from sequence information. Molecular BioSystems 2013, 9:133–42.

Wu T, Wang J, Liu C, Zhang Y, Shi B, et al: NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Research 2006, 34:D150–2.

Xue LC, Dobbs D, Honavar V: HomPPI: a class of sequence homology based protein-protein interface prediction methods. BMC Bioinformatics 2011, 12:244.

Zvereva AS, Pooggin MM: Silencing and innate immunity in plant defense against viral and non-viral pathogens. Viruses 2012, 4:2578–97.

CHAPTER 3. PREDICTING RNA-PROTEIN INTERACTIONS USING ONLY SEQUENCE INFORMATION

Modified from a paper published in BMC Bioinformatics, 2011, 12:489

Usha K Muppirala, Vasant Honavar and Drena Dobbs

Abstract

Background

RNA-protein interactions (RPIs) play important roles in a wide variety of cellular processes, ranging from transcriptional and post-transcriptional regulation of gene expression to host defense against pathogens. High throughput experiments to identify RNA-protein interactions are beginning to provide valuable information about the complexity of RNA-protein interaction networks, but are expensive and time consuming. Hence, there is a need for reliable computational methods for predicting RNA-protein interactions.

Results

We propose ***RPISeq***, a family of classifiers for predicting ***R***NA-***p***rotein ***i***nteractions using only ***seq***uence information. Given the sequences of an RNA and a protein as input, *RPIseq* predicts whether or not the RNA-protein pair interact. The RNA sequence is encoded as a normalized vector of its ribonucleotide 4-mer composition, and the protein sequence is encoded as a normalized vector of its 3-mer composition, based on a 7-letter reduced alphabet representation. Two variants of *RPISeq* are presented: *RPISeq-SVM*, which uses a Support Vector Machine (SVM) classifier and *RPISeq-RF*, which uses a Random

Forest classifier. On two non-redundant benchmark datasets extracted from the Protein-RNA Interface Database (PRIDB), *RPISeq* achieved an AUC (Area Under the Receiver Operating Characteristic (ROC) curve) of 0.96 and 0.92. On a third dataset containing only mRNA-protein interactions, the performance of *RPISeq* was competitive with that of a published method that requires information regarding many different features (e.g., mRNA half-life, GO annotations) of the putative RNA and protein partners. In addition, *RPISeq* classifiers trained using the PRIDB data correctly predicted the majority (57-99%) of non-coding RNA-protein interactions in NPInter-derived networks from *E. coli*, *S. cerevisiae*, *D. melanogaster*, *M. musculus*, and *H. sapiens*.

Conclusions

Our experiments with *RPISeq* demonstrate that RNA-protein interactions can be reliably predicted using only sequence-derived information. *RPISeq* offers an inexpensive method for computational construction of RNA-protein interaction networks, and should provide useful insights into the function of non-coding RNAs. *RPISeq* is freely available as a web-based server at http://pridb.gdcb.iastate.edu/RPISeq/.

Background

Most of the essential molecular functions of cells are governed by interactions of proteins with other proteins, nucleic acids and small ligands. Computational studies of protein interaction data have helped identify protein-protein interaction PPI networks in various organisms (Lees *et al.*, 2011, Wang *et al.*, 2011). Similarly, studies on DNA-protein interactions have allowed construction of transcription factor-gene regulatory networks (Lee

2002, Martínez-antonio, 2011). In contrast, although several ribonucleoprotein (RNP) complexes have been extensively characterized (e.g., the ribosome, the spliceosome), post-transcriptional regulatory networks that are mediated by RNA-protein interactions (RPIs) are much less well studied (Kishore *et al.*, 2010, Mittal *et al.*, 2009, Tsvetanova *et al.*, 2010, Hafner *et al.*, 2010, Hafner *et al.*, 2010). In addition to their roles in controlling gene expression at the post-transcriptional level, RPIs regulate numerous fundamental biological processes, ranging from DNA replication and transcription, to pathogen resistance, to viral replication (Hogan *et al.*, 2008, Licatalosi *et al.*, 2010, Sola *et al.*, 2011, Li *et al.*, 2011). Recently, high-throughput experiments have provided evidence for large numbers of RNA binding proteins in cells, and are beginning to identify and characterize pairs of RNAs and proteins that participate in RPIs (Baroni *et al.*, 2008, Barkan, 2009, Charon *et al.*, Kaymak *et al.*, 2010, Kim *et al.*, 2009, Pacheco *et al.*, 2010). At present, however, our understanding of RNA binding proteins lags far behind our knowledge of regulatory DNA binding proteins, such as transcription factors and replication factors.

Computational studies of RNA-protein interactions have largely focused on the "interface prediction problem", i.e., the problem of identifying the amino acid residues in a protein that are likely to bind to an RNA (Terribilini *et al.*, 2006, Pérez-Cano *et al.*, 2010, Zhou *et al.*, 2009). Only a few studies to date have focused on the "partner prediction problem", i.e., identification of specific RNA interaction partner(s) for a known RNA binding protein, or protein binding partner(s) for non-coding RNAs (ncRNAs). Although large-scale experimental analyses of RPIs such as RNAcompete (Ray *et al.*, 2009), RIP-Chip (Keene *et al.*, 2006), HITS-CLIP (Licatalosi *et al.*, 2008), PAR-CLIP (Hafner *et al.*, 2010) are now providing valuable data about networks of RNA-protein interactions, these

experiments are expensive and time-consuming. Thus, there is a compelling need for computational methods to accurately predict RPIs and to construct RNA-protein interaction networks. Given the limited number of structurally characterized RNA-protein complexes available in the PDB (Berman *et al.*, 2000) at present (1,092 as of June 13, 2011) and the current availability of only one database of ncRNA-protein interactions (NPInter (Wu *et al.*, 2006)), it would be especially valuable to develop sequence-based methods that can be used to identify potential RNA-protein partners in the absence of experimental structural information regarding either partner.

Machine learning offers one of the most cost-effective approaches to constructing predictive models in settings where experimentally validated training data are available. At present, however, it is unclear whether the available experimental data regarding RNA-protein interactions are sufficient for successfully training classifiers using machine learning algorithms. Against this background, this study explores machine learning approaches to train sequence-based classifiers for predicting RPIs.

Results

As a first step towards computational construction of RPI networks, we focused on the following question: Given the sequence of an RNA-binding protein, can we predict whether it interacts with a given RNA sequence? In developing sequence-based methods to answer this question, we considered several reduced and alternative alphabet representations of the input protein and RNA sequences. Shen *et al*. (Shen *et al.*, 2007) used a Conjoint Triad Feature (CTF) representation to successfully predict protein-protein interactions. The CTF representation essentially encodes each protein sequence using the normalized 3-gram

frequency distribution extracted from a 7-letter reduced alphabet representation of the protein sequence (See *Methods* for details). A recent study (Shao *et al.*, 2009) demonstrated the utility of the CTF representation for predicting whether a given protein is an RNA binding protein. Inspired by these studies, we chose to encode each protein sequence using the normalized *k*-gram frequency distributions extracted from the 7-letter reduced alphabet representation of the sequence. The choice of k=3 yielded the best results. We also explored several alternative representations of RNA sequences and settled on encoding each RNA sequence using normalized 4-gram frequencies extracted directly from the 4-letter ribonucleotide alphabet representation of the RNA sequence.

Our choice of Random Forest (RF) and Support Vector Machine (SVM) classifiers was motivated by several studies that have successfully used them on classification tasks that are closely related to the RPI prediction (Wang *et al.*, 2010, Hwang *et al.*, 2011, Chen *et al.*, 2005, Liu *et al.*, 2010). To rigorously evaluate the performance of these methods, we generated two non-redundant benchmark datasets, RPI2241 and RPI369, from PRIDB (Lewis *et al.*, 2011), a comprehensive database of RNA-protein complexes extracted from the PDB (Berman *et al.*, 2000). Most of the RNA-protein pairs in RPI2241 correspond to RPIs involving rRNAs or ribosomal proteins; the rest correspond to RPIs involving other ncRNAs or mRNAs. RPI369 is corresponds to RPIs extracted from non-ribosomal complexes in RPI2241. “Negative” examples of non-interacting RNA-protein pairs were generated by randomly pairing proteins with RNAs and excluding the known interacting pairs (see *Methods* for details).

RPISeq classifiers can reliably predict RNA-protein interactions

We compared the performance of *RPISeq-SVM* and *RPISeq-RF* classifiers to predict RPIs, using the benchmark datasets described above. Table 3.1 summarizes the prediction results obtained in 10-fold cross-validation experiments. On the RPI2241 dataset, the prediction accuracy was 89.6% (RF) and 87.1% (SVM); precision and recall for both classifiers was greater than 87%. On the RPI369 dataset, performance of both classifiers was considerably lower with an average accuracy of only 76.2% (RF) and 72.8% (SVM). Notably, values of the F-measure (weighted average of precision and recall) were greater than 0.70 for both classifiers on both datasets. Thus, the performance of classifiers estimated using 10-fold cross-validation on the larger RPI2241 dataset, which includes ribosomal data, is considerably better than that estimated using the RPI369 dataset, from which ribosomal data have been excluded. We also performed leave-one-out cross validation for the RF classifier. The results were not significantly different from 10-fold cross-validation experiments.

Table 3.1 Performance evaluation of RPISeq. Results of 10-fold-cross-validation experiments using RPI2241 and RPI369 datasets.

Dataset	Classifier	Accuracy %	Precision	Recall	F-measure
RPI2241	Random Forest	89.6	0.89	0.90	0.90
RPI2241	SVM	87.1	0.87	0.88	0.87
RPI369	Random Forest	76.2	0.75	0.78	0.77
RPI369	SVM	72.8	0.73	0.73	0.73

The performance statistics reported in Table 3.1 were obtained using classifiers designed to provide high prediction accuracy. By varying the classification threshold value, the prediction specificity can be increased at the expense of a decrease in sensitivity. The corresponding trade-off between true positive rate and false positive rate can be seen from the receiver operating characteristic (ROC) curve shown in Figure 3.1. Consistent with the results in Table 3.1, ROC AUCs of 0.97 (RF) and 0.92 (SVM) were obtained for predictions on the RPI2241 dataset, with lower values of 0.85 (RF) and 0.81 (SVM) on the RPI369 dataset. For both classifiers, the AUC of ROC is significantly greater than 0.50 (random), indicating the feasibility of predicting RPIs using only sequence information from the RNA and protein as input.

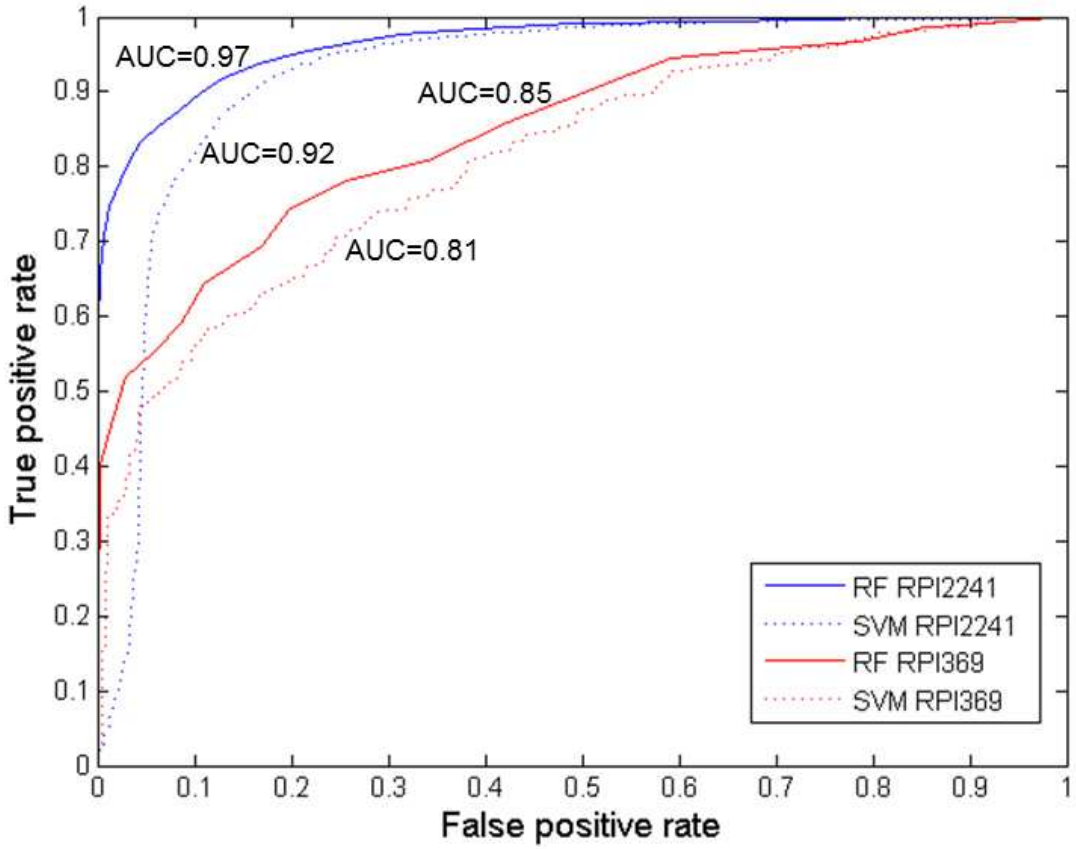


Figure 3.1 Performance of RPISeq classifiers in predicting RPIs. Receiver operating characteristic (ROC) curves for RPI predictions, illustrating the trade-off between true positive rate and false positive rate for *RPISeq-RF* (random forest) and *RPISeq-SVM* (support vector machine) classifiers, using two datasets, RPI2241 and RPI369. The area under the curve (AUC) of each ROC is shown next to the curve. The AUC for a perfect classifier is 1, and for a random classifier = 0.5.

Comparison with other methods for predicting RNA-protein interactions

Bellucci *et al*. (2011) used a variety of physicochemical properties (e.g., hydrogen-bonding propensities, secondary structure propensities) of proteins and RNAs to predict the interaction propensities for individual residues in the RNA and protein sequences of a potentially interacting pair. Because the catRAPID server [http://tartaglialab.crg.cat] does not

directly report predictions as to whether or not a specific RNA-protein pair is expected to interact (the “partner prediction problem”), we were not able to directly compare our results with their method (Bellucci *et al.*, 2011).

Pancaldi and Bähler *et al.* (2011) also employed RF and SVM classifiers, but their method uses more than 100 different features of mRNA and proteins, extracted from the literature or computed from the protein and RNA sequences to make predictions. Examples of such features include mRNA half-life, predicted protein secondary structure, Gene Ontology annotation, relative abundance of each amino acid, codon bias. Using a dataset of 5,166 positive mRNA-protein RPI partners derived from Hogan *et al.*, (2008), and 5,166 randomly generated negative examples of mRNA-protein pairs, Pancaldi and Bähler reported an average accuracy of 70% in 2-fold cross-validation tests using an RF classifier based on 500 trees, and 68% using an SVM classifier using an RBF kernel with optimized parameters (Pancaldi and Bähler, 2011). They also reported that 5-fold and leave-one-out experiments gave comparable results. We performed 10-fold cross-validation experiments on the same dataset using *RPISeq-RF*, which uses only sequence information. Our RF classifier achieved an accuracy of 68%, based on 500 trees, results comparable to the 70% reported for the RF classifier of Pancaldi and Bähler (2011). Our SVM classifier, using a normalized polykernel, gave less accurate predictions (61%) than the SVM of Pancaldi and Bähler (68%).

In the Pancaldi and Bähler study, only 5,166 out of a total of 13,243 positive mRNA-protein pairs were actually used for prediction, because some of the features required by the classifiers were not available for the remaining 8,000 pairs (Pancaldi and Bähler, 2011). When we tested our method using all 13,243 pairs for cross-validation, the prediction accuracies increased to 78% for the RF and 65% for SVM classifier. Taken together, our

experiments indicate that the sequence-based method proposed here and the multiple feature-based method of Pancaldi and Bähler have comparable performance in predicting mRNA-protein interactions. Further, our results suggest that sequences of mRNAs and proteins carry sufficient information to allow reasonable predictions regarding whether or not a given mRNA and protein interact. Because feature information required by the method of Pancaldi and Bähler may not be available in many cases, our proposed method complements theirs, and may be more generally applicable for predicting ncRNA-protein partners, in addition to mRNA-protein partners.

Predicting ncRNA-protein interaction networks

An important potential application of ***RPISeq*** is computational construction of RNA-protein interaction networks. Recently, Nacher and Araki (2010) used RPIs from the NPInter database (Wu *et al.*, 2006), a database of non-coding RNA-protein interactions, to construct non-coding RNA-protein networks for several different model organisms. Their study revealed significant similarities between ncRNA-protein and transcription factor-gene regulatory networks. To explore whether *RPISeq* could be useful for constructing networks of ncRNA-protein interactions, we evaluated our method in predicting RPIs in networks derived from NPInter. Because the NPInter RPI pairs do not include any pairs derived from ribosomes, in this experiment, we also compared the performance of models trained on the RPI369 (which lacks ribosomal sequences) versus RPI2241, to evaluate the potential effect of strong ribosomal sequence bias on performance.

Table 3.2 RPISeq predictions on NPInter dataset using RF and SVM classifiers trained on RPI2241.

Organism	Total RPI pairs	Pairs predicted by RF (%)	Pairs predicted by SVM (%)
H. sapiens	1189	888 (74.7)	681 (57.3)
S. cerevisiae	254	249 (98.0)	252 (99.2)
M. musculus	120	98 (81.7)	85 (70.8)
D. melanogaster	81	80 (98.8)	72 (88.9)
E. coli	37	34 (91.9)	25 (67.6)
Total	**1681**	**1349 (80.2)**	**1115 (66.3)**

Table 3.3 RPISeq predictions on NPInter dataset using RF and SVM classifiers trained on RPI369.

Organism	Total RPI pairs	Pairs predicted by RF (%)	Pairs predicted by SVM (%)
H. sapiens	1189	808 (68.0)	988 (83.1)
S. cerevisiae	254	168 (66.1)	226 (89.0)
M. musculus	120	81 (67.5)	111 (92.5)
D. melanogaster	81	38 (46.9)	53 (65.4)
E. coli	37	20 (54.0)	24 (64.9)
Total	**1681**	**1115 (66.3)**	**1402 (83.4)**

Tables 3.2 and 3.3 show the number of RPI pairs correctly predicted for each organism. When trained on the RPI2241 dataset (Table 3.2), the RF classifier correctly predicted ~ 80% (1,349 of 1,681 total interactions). The output probabilities of *RPISeq* are estimates of interaction propensities for a specific RNA-protein pair. In Tables 3.2 and 3.3,

the probability threshold used for "positive" interactions was 0.50. Among the 1,349 interactions predicted by the RF classifier, only 119 were predicted with probabilities ≥ 0.80, and another 1,230 interactions were predicted with probabilities in the range 0.50-0.80. The SVM classifier generally had slightly lower performance, correctly predicting ~ 66% of the interactions.

In contrast, when trained on the RPI369 dataset, the SVM classifiers out-performed the RF classifiers (Table 3.3). Overall, the SVM classifier correctly predicted 1,402 (83%) and the RF classifier correctly predicted 1,115 (66%) of the interactions. Among the 1,402 interactions correctly predicted by SVM classifier, more than 850 interactions were predicted with probabilities ≥ 0.80, and another 525 interactions were predicted with probabilities in the range 0.50 to 0.80. For the RF classifier, only 50 interactions were predicted with probabilities ≥ 0.80.

With regard to the effects of ribosomal sequence bias, these results are somewhat difficult to interpret. The best "overall" prediction performance was obtained using the SVM classifier trained on the RPI369 dataset, with 83.4% interactions correctly predicted; the RF classifier trained on the RPI2241 dataset correctly predicted 80.2% of the total interactions. This difference in overall performance, based on the combined data from all five organisms, is relatively small. In contrast, however, differences in performance of classifiers trained on the two different datasets are much larger when predictions for each model organism are considered individually. For example, for *D. melanogaster*, substantially better predictions were obtained with an RF classifier trained on the RPI2241 dataset (98.8%) versus an RF classifier trained on the RPI369 dataset (46.9%). In contrast, for predicting human and mouse RNA-protein interactions, SVM classifiers trained on the RPI369 dataset (which excludes the

ribosomal sequences) provide the best prediction performance. For yeast RPIs, both the RF and SVM classifiers trained on RPI2241 generated excellent predictions, 98.0% and 99.2%, respectively, whereas classifiers trained on RPI369 made more errors, with correct predictions for 66.1% (RF) and 89.0% (SVM) of the cases.

Figure 3.2 shows the ncRNA-protein interaction network from *S. cerevisiae,* based on the data in NPInter. In Figure 3.2A, *RPISeq* predictions obtained using classifiers trained on the RPI2241 dataset are mapped onto the network. As described above, the SVM classifier (right) makes more correct predictions (green edges) and fewer incorrect predictions, i.e., false negatives, (red edges) than the RF classifier (left). In Figure 3.2B, RPI*Seq* predictions made using classifiers trained on the RPI369 dataset, which results in more errors, are shown.

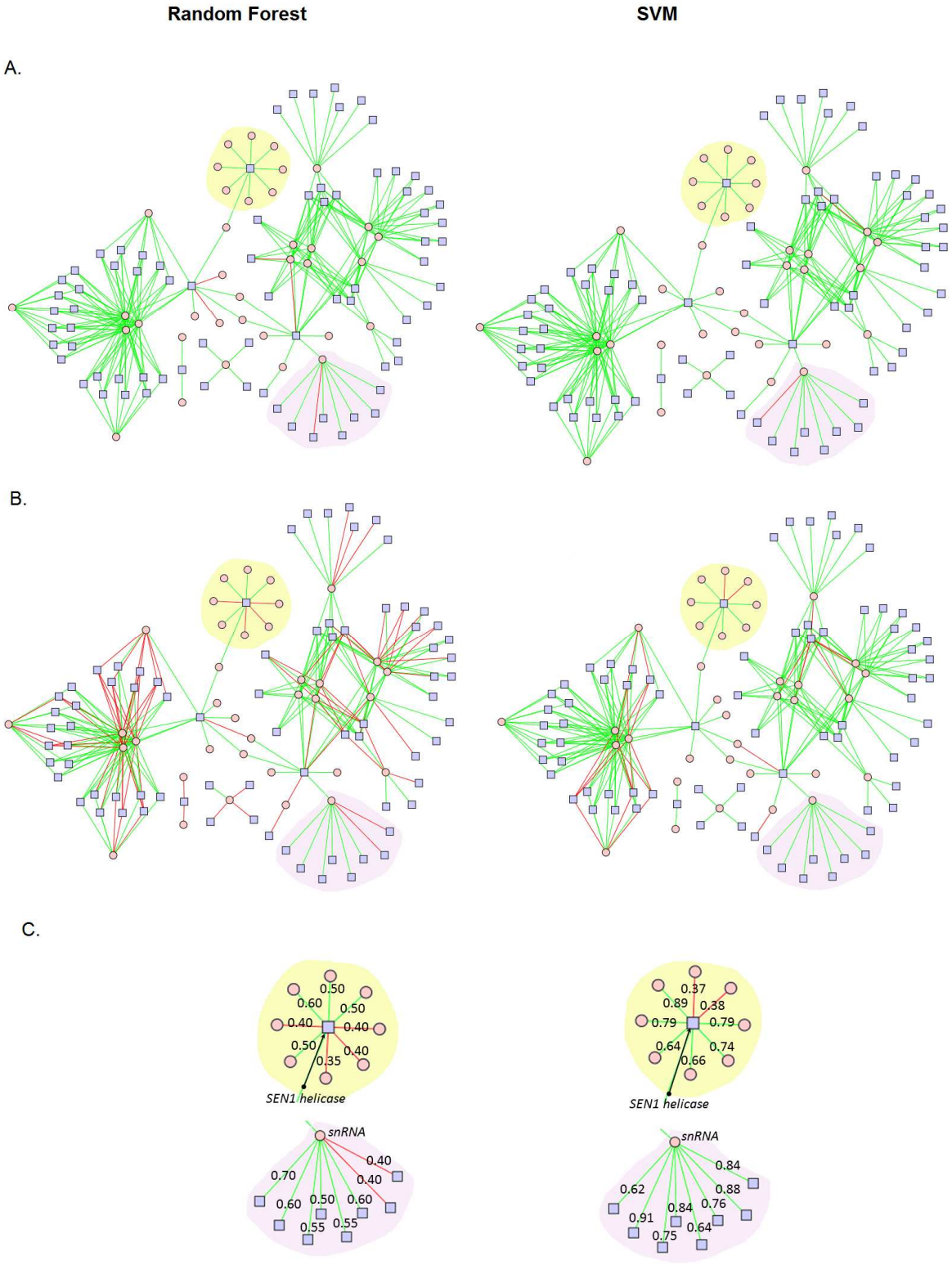


Figure 3.2 RPISeq correctly predicts a majority of interactions in a yeast ncRNA-protein

interaction network. Circles represent RNA and squares represent proteins. RNA-protein pairs predicted by RPISeq-RF or RPISeq-SVM classifiers are mapped onto the network of validated interactions, with correctly predicted interactions shown as green edges and incorrect predictions as red edges. (A) Predicted interactions using classifiers trained on the RPI2241 dataset. Among 254 known interactions, RPISeq-RF and RPISeq-SVM classifiers correctly predicted all except 5 and 2 edges, respectively. A protein hub, highlighted in yellow, shows interactions of a helicase (SEN1) with several snoRNAs. One of several RNA hubs, highlighted in purple, illustrates interactions of an snRNA (u4560) with various Sm-like proteins in the LSM complex. (B) Predicted interactions using classifiers trained on RPI369 dataset. Among 254 known interactions, RPISeq-RF classifier correctly predicted 168 (66%) and RPISeq-SVM correctly predicted 226 (89%). A protein hub highlighted in yellow, shows interactions of a helicase (SEN1) with 8 snoRNAs. One of several RNA hubs, highlighted in purple, illustrates interactions of an snRNA (u4560) with various Sm-like proteins in the LSM complex. (C) An enlarged view of the protein (SEN1) and RNA (snRNA) hubs described in B. above. Edges are labelled with the interaction probabilities predicted by RPISeq-RF (left) and RPISeq-SVM (right) classifiers, providing estimates of the relative pairwise interaction propensities.

One protein hub (highlighted in yellow), which appears as a green square node with connections to several RNA nodes (pink circles), is apparent in these views of the network. It corresponds to the yeast SEN-1 helicase, which is known to interact with several snoRNAs (Ursic *et al.*, 2004). Several RNA hubs, represented by red circular nodes, each connected to several green protein nodes, are also apparent. One of these RNA hubs (highlighted in purple), corresponds to snRNA u4560, which interacts with various Sm-like proteins in the LSM complex (Vidal *et al.*, 1999).

Figure 3.2C shows an enlarged view of these hubs, extracted from Figure 3.2B. Edges are labelled with the interaction probabilities predicted by each classifier. Using classifiers trained on the RPI369 dataset, the RF classifier made more errors (i.e., predicted a known

interaction with probability < 0.5) than the SVM classifier in both cases: for SEN-1 helicase, the RF classifier correctly identified only 4 out of 8 known snoRNA interactions, whereas the SVM classifier correctly identified 6 out of 8. Similarly, of 8 proteins known to interact with snRNA u4560 in yeast, the RF classifier identified 6, while the SVM classifier correctly identified all 8 interaction partners. Notably, as shown in Figure 3.2A, both RF and SVM classifiers trained on the RPI2241 dataset correctly identified all 8 RNA interaction partners of the SEN-1 helicase, and both classifiers missed only 1 of 8 protein interaction partners of the snRNA u4560.

Discussion

Regulation of gene expression at the post-transcriptional level is often mediated by interactions between RNA binding proteins and mRNAs or ncRNAs (Kishore *et al.*, 2010, Licatalosi *et al.*, 2010, Blencowe *et al.*, 2009). In this work, we present a new method, *RPISeq*, for predicting RNA-protein interaction partners, using only sequence information, with up to 90% average accuracy. We also demonstrate, that *RPISeq* can effectively predict RNA-protein interaction networks, based on evaluation using available data from five model organisms.

Sequence-based prediction of RNA-protein interactions

While several computational methods for predicting networks of protein-protein interactions have been developed (Lees *et al.*, 2011, Wang *et al.*, 2011), very few studies have focused on computational analysis or prediction of RNA-protein interactions (Lee *et al.*, 2002, Martínez-antonio, 2011). One of the major challenges in solving the “partner

prediction problem" for RNA-protein interactions is the limited amount of experimental data currently available. Unlike the "interface prediction problem," for which detailed structural information for more than 1,000 RNA-protein complexes is available in the PDB, mRNA partners for only a handful of RBPs are known (Hogan *et al.*, 2008). Currently, two basic types of information regarding RNA-protein interaction partners are widely available: i) experimentally-determined structures of RNA-protein complexes, available in primary resources such as the PDB (Berman *et al.*, 2000) and NDB (Berman *et al.*, 1992), and secondary resources such as PRIDB (Lewis *et al.*, 2011) and BIPA (Lee *et al.*, 2009); and ii) experimental data from *in vivo* or *in vitro* cross-linking studies focused on individual proteins (e.g., SFRS1 (Sanford *et al.*, 2009), PUF (Gerber *et al.*, 2004) or from high throughput RNA-binding microarrays (Ray *et al.*, 2009), stored in repositories such as NPInter (Wu *et al.*, 2006), CLIPZ (Khorshid *et al.*, 2010) and RBPDB (Cook *et al.*, 2010).

RPISeq requires only sequence information to generate predictions. In the current version of *RPISeq*, the classifiers were trained using only RPIs for which experimental structures are available. RPI2241 is a non-redundant training dataset consisting of 2241 interacting RNA-protein pairs, and includes a wide variety of different functional classes of proteins and RNA (e.g., rRNA, tRNA, miRNA, mRNA). rRNA-ribosomal protein pairs constitute ~ 40% of the total, reflecting the predominance of ribosomal structures in the current version of the PDB. To investigate the impact of this bias on machine learning methods for predicting RPIs, we also generated a smaller dataset of 369 RNA-protein partners (RPI369), from which all rRNA-containing complexes had been removed (see *Methods* for details).

We used RPI2241 and RPI369 as non-redundant benchmark datasets for developing and rigorously evaluating the performance of various machine learning classifiers. In cross-validation experiments, classifiers trained and tested on the larger dataset had superior prediction performance, indicating that the greater number and diversity of complexes in RPI2241, relative to RPI369, has a stronger positive effect on classification accuracy than the potentially negative effect of sequence bias in RPI2241. When we evaluated classifiers using independent datasets of RPIs from NPInter, however, classifiers trained on RPI369, in some cases, had better prediction performance. The basis for this observation is currently under investigation.

To identify sequence features of the proteins and RNA important in determining their specific interactions, we analyzed the features most frequently used by the Random Forest classifier to predict interacting partners (see *Methods* for details).

The four most often selected RNA tetrads were: *AUUC, AGUG, UUUU, UCAA*. Notably, these tetrads were found in the interfacial region in only 15% of the cases examined. The most frequently selected conjoint triad in protein sequences was {*I, L, F, P*}{*A, G, V*}{*R, K*}, which represents twenty-four possible amino acid triplets (e.g., *IAR, IAK, IGR, IGK*...). The complete list of important RNA and protein features is provided in Supplemental Data S1. Although additional experiments and analyses of these features will be required to extract precise “rules” that specify a particular RNA-protein interaction, our current analysis indicates that at least 50 features (a combination of RNA and protein features) are required to accurately classify a given RNA-protein pair as interacting or not.

In this study, *RPISeq* accurately predicted RPIs in both cross-validation experiments using the benchmark datasets and in experiments on independent datasets. This suggests that

normalized *k*-mer frequency distributions of RNA and protein sequences (specifically, reduced alphabet representations of protein sequences) in combination with appropriate machine learning methods, provide an effective approach to construct RPI predictors. Because the data used in this study represent only a small fraction of cellular RNA-protein complexes and interactions, we anticipate that more accurate predictions will be possible when larger and more diverse datasets of experimentally validated RPIs become available.

Comparison with other available methods

The method of Pancaldi and Bähler (2011), which was developed to predict mRNA-protein interactions (rather than ncRNA-protein interactions), also uses RF and SVM classifiers, but requires a much more extensive set of features regarding the mRNAs and proteins. Input for the classifiers, which consists of a vector constructed by concatenating the features of potential RNA and protein partners (e.g., isoelectric point of protein, protein localization, mRNA half-life), cannot be extracted or calculated from sequence information alone. This requirement restricts the applicability of this method in practice: Pancaldi and Bähler were not able to extract the necessary features for a majority of interactions in their RPI dataset. The *RPISeq* methods do not suffer from this limitation because they require only sequence-derived features to make reliable predictions. In fact, the performance of *RPISeq* improved substantially (by 8% in accuracy) when evaluated on the entire dataset of Pancaldi and Bähler. Thus, for predicting mRNA-protein interactions, the sequence-based approach implemented in *RPISeq* provides performance comparable to that of classifiers that require a more extensive set of features, including those that cannot be extracted from RNA and protein sequences alone.

Application of *RPISeq* to constructing RNA-protein interaction networks

Encouraged by the success of *RPISeq* in predicting specific RPIs, we examined its effectiveness in constructing RNA-protein interaction networks in several model organisms, using only information derived from RNA and protein sequences. The networks were extracted from the "ncRNA binds protein" category of NPInter [27], currently the only available database of functional interactions of ncRNA with proteins. *RPISeq* was able to successfully predict the interactions of a single protein with multiple RNAs (protein hubs), as well as interactions of a single RNA with multiple proteins (RNA hubs).

In the case of the yeast, *S. cerevisiae*, *RPISeq* provided excellent predictions of RPIs: both the RF and SVM classifiers trained on the RPI2241 dataset correctly predicted > 98% of interactions in the NPInter database (Wu *et al.*, 2006). The *RPISeq-RF* classifier trained on the RPI2241 dataset also correctly identified a large majority of interactions in the *D. melanogaster* (99%) and *E. coli* (92%) networks. For human and mouse networks, however, classifiers trained on the RPI369 dataset gave better performance, with the *RPISeq-SVM* classifier correctly identifying 83% of the interactions in human and 93% in the mouse. It is important to note that these evaluations are based on predicting only known *"positive"* interactions currently available in NPInter (Wu *et al.*, 2006); *"negative"* data regarding non-interacting protein-RNA-protein pairs are not included in NPInter. Because the experimental data in NPInter are incomplete, it is problematic to assume that RNA-protein pairs not included in NPInter do not, in fact, interact. Also, some experimentally-determined RPIs included in NPInter could correspond to false positives.

Given the relatively small sizes of the RNA-protein networks analyzed in this study, differences in the results obtained using different classifiers to predict RPIs in different

species must be interpreted with caution. It will be important to evaluate these methods on larger, more complete datasets of experimentally validated RNA-protein interactions as they become available. On the whole, our results suggest that *RPISeq* should be valuable for constructing and analyzing regulatory RNA-protein interaction networks.

Conclusion

In this work, we tested whether *RPISeq*, a family of purely sequence-based classifiers, can be used to predict whether a specific RNA-protein pair is likely to interact. Our results demonstrate that the corresponding RNA and protein sequences alone contain sufficient information to allow reliable prediction of RPIs. Such predictions can be used to: (i) identify putative RNA partners of a target protein, or protein partners of a target RNA; and (ii) computationally construct RNA-protein interaction networks. The datasets used in this study are relatively small compared with the large number of RNA-protein complexes and diverse interactions that occur in cells. The increasing availability of transcriptome-wide experimental data should lead to improvements in computational methods for predicting RNA-protein interactions and for modelling regulatory networks of RNA-protein interactions. *RPISeq* is freely available as a web-based server at http://pridb.gdcb.iastate.edu/RPISeq/.

Methods

RPI benchmark datasets derived from structure-based experimental data

For training and testing classifiers, two benchmark non-redundant datasets of RNA-protein interacting pairs were extracted from 943 protein-RNA complexes in PRIDB using an

8 Å distance cut-off (Lewis *et al.*, 2011). PRIDB is a database of protein-RNA interfaces calculated from protein-RNA complexes in the PDB (Berman *et al.*, 2000). The original 943 complexes from PRIDB contained a total of 9,689 protein chains and 2,074 RNA chains; the final dataset RPI2241 (see below), which contains a total of 952 protein chains and 443 RNA chains, was derived from these complexes by applying the following criteria. Redundant protein sequences (i.e., with ≥ 30% sequence identity) interacting with similar RNA sequences (i.e., with ≥ 30% sequence identity) were discarded. Also, redundant RNA sequences (i.e., with ≥ 30% sequence identity) interacting with similar protein sequences (i.e., with ≥ 30% sequence identity) were discarded. Only proteins whose length is greater than 25 and RNAs at least 15 nucleotides long were retained. This resulted in a dataset of "positive" examples, RPI2241, consisting of 2241 experimentally validated RNA-protein pairs (Supplemental Data S2).

To generate a balanced dataset of "non-interacting RNA-protein pairs" (negative examples), we randomly paired the RNAs and proteins from the 943 protein-RNA complexes and removed similar interacting RNA-protein pairs (a randomly generated pair A-B was discarded if there exists a positive interaction pair C-B, and A and C share ≥ 30% sequence identity). Because ~40% of RNA-protein complexes in the PDB correspond to ribosomal structures, the RPI2241 dataset is also strongly biased towards ribosomal RPIs. Thus, we constructed a second dataset, RPI369, which is a subset of RPI2241 generated by removing all RPIs that contain ribosomal proteins or ribosomal RNAs (Supplementary Data S3). RPI369 contains only non-ribosomal complexes (e.g., tRNA, mRNA, viral RNA, miRNA).

RPI benchmark datasets derived from non-structure-based experimental data

For evaluation of our method on independent RPI datasets, we used two datasets of RPIs obtained from RNA immunoaffinity purification and microarray experiments, published by Hogan *et al* (2008). One dataset comprises 5,166 mRNA-protein interactions; this dataset was also used in the study of Pancaldi and Bähler (2011). The second dataset is larger, consisting of 13,243 RPIs, and including all 5,166 interactions in the smaller dataset. Pancaldi and Bähler were not able to evaluate their method on this larger dataset because of missing feature information for RNAs and proteins involved in these interactions. Because *RPISeq* uses only sequence information, we were able to evaluate our method using all of the available data.

To test the ability of *RPISeq to* predict ncRNA-protein interaction networks, we used the NPInter database (http://www.panrna.org/NPInter/), which includes eight different categories of functional interactions between non-coding RNAs, but excludes ribosomal RNAs and proteins. We extracted only those interactions for which there is experimental evidence for physical association of ncRNA with a protein, i.e. the 'ncRNA binds protein' category.

Alternative representations of protein and RNA sequences

Each RNA-protein pair is represented as a 599-feature vector, in which 343 features are used to encode the protein sequence and 256 features are used to encode the RNA sequence. Proteins are encoded using the conjoint triad feature (CTF) representation previously used by Shen *et al* (2007). In this method, the 20 amino acids are classified into 7 groups according to their dipole moments and the volume of their side chains: {*A, G, V*}, {*I,*

L, F, P}, {*Y, M, T, S*}, {*H, N, Q, W*}, {*R, K*}, {*D, E*}, {*C*}. Each protein sequence is then encoded using the 7-letter reduced alphabet. Each protein feature represents the normalized frequency of the corresponding conjoint triad, i.e., 3-mer in the 7-letter reduced alphabet representation of the protein sequence. Thus, each protein sequence is represented by a 343 (7×7×7) dimensional vector, where each element of the vector corresponds to the normalized frequency of the corresponding 3-mer in the sequence (see (Shen *et al.*, 2007) for details). Based on results of preliminary tests comparing the normalized *k*-mer frequency representation of RNA sequences for different values of *k*, we chose to encode RNA sequences using a 4×4×4×4 or 256-dimensional vector, in which each feature represents the normalized frequency of the corresponding 4-mer appearing in the RNA sequence (e.g., *AAUG, CGAU, GGCC).*

Machine learning Algorithms

The support vector machine (SVM), a machine learning algorithm developed by Vapnik [47], is widely used for classification and regression tasks. SVM is a binary classification method that takes two differently labeled classes as input and outputs a model to classify unlabeled data. SVM maps the input onto a higher dimensional space and constructs a hyperplane to separate the two classes with a maximum margin. In this work, we used the SMO implementation in Weka 3.7 [48]. The SMO classifier implements the sequential minimal optimization algorithm to train SVMs. We used a normalized polykernel function with = 1.0E-12 and C=1.0 and built logistic models on the SVM to output probability estimates for the predictions. The normalized polykernel gave the better performance than other kernels tested, including the RBF kernel (data not shown).

Random Forest (RF) methods have been successfully applied to many problems in bioinformatics, including prediction of protein-protein interactions [32, 33, 49]. Random Forest is an ensemble classifier consisting of many tree-structures classifiers. For the problem addressed here, in which the number of input features is large, significant improvements can be expected by employing feature selection [50]. We used the Random Forest implementation in Weka 3.7 for model building and evaluation. We constructed the RF classifiers with 20 trees (unless otherwise indicated) and 10 features were evaluated at each node. For performing feature selection, we used *AttributeSelection* class in Weka toolkit. We used *wrapper subset evaluator* in combination with Random Forest classifier and best first search method.

Performance Evaluation

Standard 10-fold cross-validation procedures were used to evaluate and compare classifier performance on the benchmark datasets. For the RF classifier, we also performed leave-one-out cross-validation; results were not significantly different from those obtained using 10-fold cross-validation (data not shown).

We computed the following statistics to measure the performance of the classifiers.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F\text{-}Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives.

The F-Measure is a composite indicator of performance that attempts to "balance" precision and recall. F-Measure values range from 0 to 1, with values close to 1 indicating better performance. The area under the curve (AUC) of the receiver operating characteristic curve (ROC) was also computed. AUC values also range from 0 to 1: the AUC = 1 for a perfect classifier and for a random classifier = 0.5.

Authors' contributions

UKM conceived the study (with DD and VGH), carried out the experiments, implemented the *RPISeq* webserver and prepared the initial draft of the manuscript. DD and VGH contributed to the experimental design, supervised the work, and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements and Funding

We thank Benjamin Lewis, Pete Zaback and Rasna Walia for valuable suggestions and comments on the manuscript. We also thank Yasser EL-Manzalawy for critical reading of the manuscript and other members of the Honavar research group for interesting discussions. The work of Vasant Honavar while working at the National Science Foundation was supported by the National Science Foundation. Any opinion, finding, and conclusions contained in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work was partially supported by funding from National Institutes of Health (GM066387 to VGH and DD) and Iowa State University's

Center for Integrated Animal Genomics (to UKM and DD). Partial funding for open access charges was provided by Iowa State University.

Additional files

The supplementary data described below is available online at this URL: http://www.biomedcentral.com/1471-2105/12/489/additional.

Supplemental Data S1 – List of RNA and protein features important for distinguishing interacting and non-interacting pairs.

Supplemental Data S2 – Positive RPIs in the RPI2241 dataset. This is a tab-delimited file with two columns. The first column is a list of proteins and the second column is a list of corresponding RNAs.

Supplemental Data S3 – Positive RPIs in the RPI369 dataset. This is a tab-delimited file with two columns. The first column is a list of proteins and the second column is a list of corresponding RNAs.

References

Barkan A: Genome-wide analysis of RNA-protein interactions in plants. Methods Mol Biol 2009, 553:13-37.

Baroni TE, Chittur SV, George AD, Tenenbaum SA: Advances in RIP-chip analysis : RNA-binding protein immunoprecipitation-microarray profiling. Methods Mol Biol 2008, 419:93-108.

Bellucci M, Agostini F, Masin M, Tartaglia GG: Predicting protein associations with long noncoding RNAs. Nature Methods 2011, 8:444-445.

Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh S-H, Srinivasan AR, Schneider B: A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J 1992, 63:751-759.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. Nucleic Acids Res 2000, 28:235-42.

Blencowe B, Brenner S, Hughes T, Morris Q: Post-transcriptional gene regulation: RNA-protein interactions, RNA processing, mRNA stability and localization. Pac Symp Biocomput 2009:545-548.

Charon C, Moreno AB, Bardou F, Crespi M: Non-protein-coding RNAs and their interacting RNA-binding proteins in the plant cell nucleus. Mol Plant 2010, 3:729-739.

Chen X-W, Liu M: Prediction of protein-protein interactions using random decision forest framework. Bioinformatics 2005, 21:4394-400.

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR: RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 2010, 39:301-308.

Gerber AP, Herschlag D, Brown PO: Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. PLoS Biol 2004, 2:E79.

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T: Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010, 141:129-141.

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T: PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. J Vis Exp 2010.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: The WEKA data mining software: An update. SIGKDD Explorations 2009, 11:10-18.

Ho TK: The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 1998, 20:832-844.

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO: Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol 2008, 6:e255.

Hwang H, Vreven T, Whitfield TW, Wiehe K, Weng Z: A machine learning approach for the prediction of protein surface loop flexibility. Proteins: Struct. Funct. Bioinf. 2011, 79:doi: 10.1002/prot.23070.

Kaymak E, Wee LM, Ryder SP: Structure and function of nematode RNA-binding proteins. Curr Opin Struct Biol 2010, 20:305-312.

Keene JD, Komisarow JM, Friedersdorf MB: RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nature protoc 2006, 1:302-7.

Khorshid M, Rodak C, Zavolan M: CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res 2010, 39:245-252.

Kim MY, Hur J, Jeong S: Emerging roles of RNA and RNA-binding protein network in cancer cells. BMB Rep 2009, 42:125-130.

Kishore S, Luber S, Zavolan M: Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. Brief Funct Genomics 2010, 9:391-404.

Lee S, Blundell T: BIPA: a database for protein-nucleic acid interaction in 3D structures. Bioinformatics 2009, 25:1559-1560.

Lee TI: Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 2002, 298:799-804.

Lees JG, Heriche JK, Morilla I, Ranea JA, Orengo CA: Systematic computational prediction of protein interaction networks. Phys Biol 2011, 8:035008.

Lewis BA, Walia RR, Terribilini M, Feguson J, Zheng C, Honavar V, Dobbs D: PRIDB: a Protein-RNA Interface Database. Nucleic Acids Res 2011, 39:D277-82.

Li Z, Nagy PD: Diverse roles of host RNA binding proteins in RNA virus replication. RNA Biol 2011, 8:305-315.

Licatalosi DD, Darnell RB: RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 2010, 11:75-87.

Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer, Blume JE, Wang X, Darnell JC, Darnell RB: HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 2008, 456:464-9.

Liu Z-P, Wu L-Y, Wang Y, Zhang X-S, Chen L: Prediction of protein-RNA binding sites by a random forest method with combined features. Bioinformatics 2010, 26:1616-1622.

Martínez-antonio A: Escherichia coli transcriptional regulatory network. Netw Biol 2011, 1:21-33.

Mittal N, Roy N, Babu MM, Janga SC: Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. Proc Natl Acad Sci U S A 2009, 106:20300-20305.

Nacher JC, Araki N: Structural characterization and modeling of ncRNA-protein interactions. Biosystems 2010, 101:10-9.

Pacheco A, Martinez-Salas E: Insights into the biology of IRES elements through riboproteomic approaches. J Biomed Biotechnol 2010, doi:10.1155/2010/458927.

Pancaldi V, Bähler J: In silico characterization and prediction of global protein-mRNA interactions in yeast. Nucleic Acids Res 2011:1-11.

Pérez-Cano L, Fernández-Recio J: Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. Proteins 2010, 78:25-35.

Ray D, Kazan H, Chan ET, Castillo LP, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR:. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nature Biotechnol 2009, 27:667-70.

Sanford JR, Wang X, Mort M, VanDyun N, Cooper DN, Mooney SD, Edenburg HJ, Liu Y: Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res 2009, 19:381-94.

Shao X, Tian Y, Wu L, Wang Y, Jing L, Deng N: Predicting DNA- and RNA-binding proteins from sequences with kernel methods. J Theor Biol 2009, 258:289-293.

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H: Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci U S A 2007, 104:4337-41.

Sola I, Mateos-Gomez PA, Almazan F, Zuñiga S, Enjuanes L: RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. RNA Biol 2011, 8:237-248.

Terribilini M, Lee J-H, Yan C, Jerniga RL, Honavar V, Dobbs D: Prediction of RNA binding sites in proteins from amino acid sequence. RNA 2006, 12:1450-62.

Tsvetanova NG, Klass DM, Salzman J, Brown PO: Proteome-wide search reveals unexpected RNA-binding proteins in Saccharomyces cerevisiae. PLoS One 2010, 5:e12671.

Ursic D, Chinchilla KJSF, Culbertson MR: Multiple protein/protein and protein/RNA interactions suggest roles for yeast DNA/RNA helicase Sen1p in transcription, transcription-coupled DNA. Nucleic Acids Res 2004, 32:2441-2452.

Vapnik V: The Nature of Statistical Learning Theory. New York: Springer; 1995.

Vidal VP, Verdone L, Mayes AE, Beggs JD: Characterization of U6 snRNA-protein interactions. RNA 1999, 5:1470-81.

Wang T-Y, He F, Hu Q-W, Zhang Z: A predicted protein-protein interaction network of the filamentous fungus Neurospora crassa. Mol Biosyst 2011.

Wang Y, Wang J, Yang Z, Deng N: Sequence-based protein-protein interaction prediction via support vector machine. J Syst Sci Complex 2010, 23:1012-1023.

Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, Sun X: Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics 2009, 25:30-5.

Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, Zhang Z, Skogerbø G, Chen L, Lu H, Zhao Y, Chen R: NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Res 2006, 34:D150-2.

Zhou P, Zou J, Tian F, Shang Z: Geometric similarity between protein–RNA interfaces. J Comput Chem 2009, 30:2738–2751.

CHAPTER 4. RPISEQ & RPINTDB: TOOLS FOR PREDICTING RNA-PROTEIN INTERACTIONS

Abstract

RPISeq 2.0 is an enhanced web-based implementation of a novel algorithm that predicts RNA-protein interaction partners. It takes a protein sequence and an RNA sequence as input and predicts the probability that the given protein and RNA physically interact with each another. The server allows for submission of multiple protein or RNA sequences, allowing users to scan a defined proteome or transcriptome for potential interacting partners. The server also allows users to query a protein sequence of interest against the RNA-Protein Interaction DataBase (RPIntDB) to identify homologous proteins and their interacting partners. RPISeq 2.0 is available at http://pridb.gdcb.iastate.edu/RPISeq.

Introduction

RNA-protein interactions (RPIs) play important roles in a wide variety of cellular processes. High throughput experiments designed to identify RNA-protein interactions are beginning to provide valuable information about the complexity of RNA-protein interaction networks, but are still expensive and time consuming. We developed **RPISeq** (Muppirala *et al.*, 2011) to address the need for reliable computational methods for predicting RNA-protein interaction partners. Whereas many computational methods and several webservers (Walia *et al.*, 2012) are available for predicting *RNA binding residues* in proteins, only five methods have been published for predicting *RNA-protein interaction partners* (Pancaldi & Bähler, 2011, Bellucci *et al.*, 2011, Muppirala *et al.*, 2011, Wang *et al.*, 2013, Lu *et al.*, 2013)

reviewed in Muppirala *et al.* (2013), and only two webservers that implement such methods are currently available: **RPISeq** (http://pridb.gdcb.iastate.edu/RPISeq/) and **catRAPID** (http://service.tartaglialab.com/page/catrapid_group).

Method

The basic RPISeq algorithm (Muppirala *et al.*, 2011) uses either Random Forest (RF) or Support Vector Machine (SVM) classifiers to predict whether a given pair of protein and RNA sequences is likely to interact, using only sequence information as input. In this approach, the protein sequence is encoded as a 343-dimensional vector, using a conjoint triad feature (CTF) method (Shen *et al.*, 2007), in which each feature represents the normalized frequency of the corresponding conjoint triads in the sequence. Similarly, the RNA sequence is encoded as a 256-feature vector, in which each feature represents the normalized frequency of the corresponding RNA tetrads. RPISeq was trained and tested on a dataset of 2,241 experimentally validated physical interactions obtained from the Protein RNA Interface Database (PRIDB) (Lewis *et al.*, 2011) and 2,241 randomly generated negative examples and achieved accuracies of 89.6% (RF classifier) and 87.1% (SVM classifier) with AUC of 0.97 (RF classifier) and 0.92 (SVM classifier) (Muppirala *et al.*, 2011). When tested on an independent dataset of 126 positive interactions generated from the NPInter (Wu *et al.*, 2006), 112 interactions were correctly predicted by RF classifier. On an independent dataset of 332 negative interactions generated by pairing known non-RNA binding proteins with RNAs in the training set, 196 (59%) were correctly predicted as negatives. All the proteins and RNAs used in these independent test sets are unique and do not overlap with the training data used to generate the models.

RPISeq webserver output

The output of the basic RPISeq algorithm is a set of probability scores (from both RF and SVM classifiers) that indicate the likelihood of interaction between the given protein and RNA pair. RNA-protein pairs with scores greater than 0.5 are predicted to interact.

A high-demand application of RPISeq is screening a large number of RNA or protein sequences for potential interaction partners. The updated version of RPISeq described here, **RPISeq (v 2.0),** provides a mechanism for accepting multiple sequences in batch mode. If the user is interested in identifying many potential RNA partners for a particular protein, the user can input the protein sequence of interest and upload a set of potential RNA target sequences as a single file in FASTA format. Similarly, a user can input one RNA sequence and upload multiple protein sequences at once. In batch submission mode, the results can be viewed online or downloaded in a tab-delimited file. Figure 4.1 A shows sample output of a single protein-RNA interaction prediction. Results of a sample batch submission are shown in Figure 4.1 B.

In summary, RPISeq (v 2.0) allows users to address 3 types of questions: 1) Does a specific protein sequence interact with a specific RNA sequence? 2) For a given protein of interest, what are its likely RNA interaction partners? 3) For a given RNA of interest, what are its likely protein interaction partners?

An advantage of RPISeq over other available methods (Pancaldi and Bähler, 2011, Bellucci *et al.*, 2011, Wang *et al.*, 2013) is its speed: RPISeq can process a single query in less than one second.

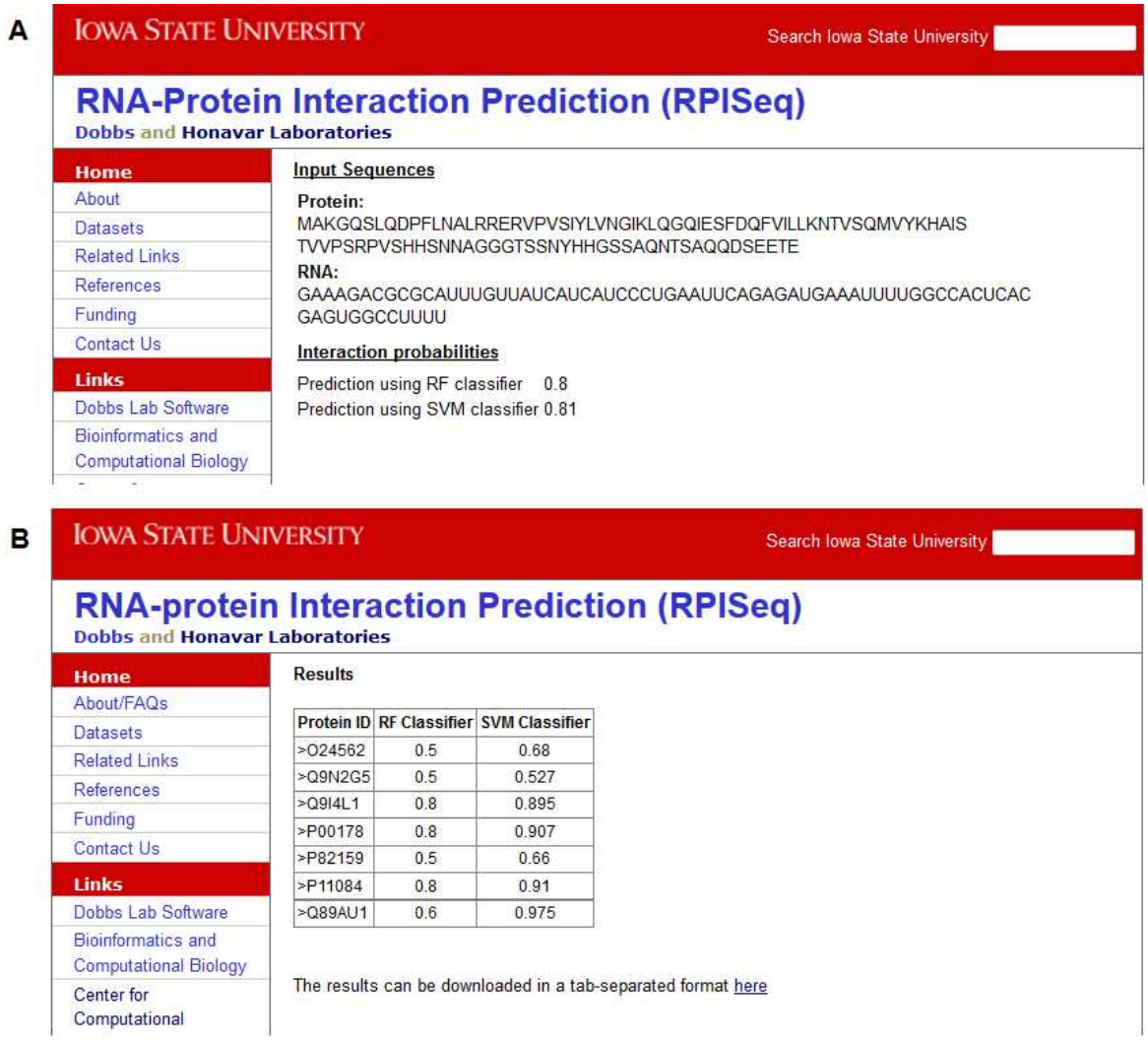
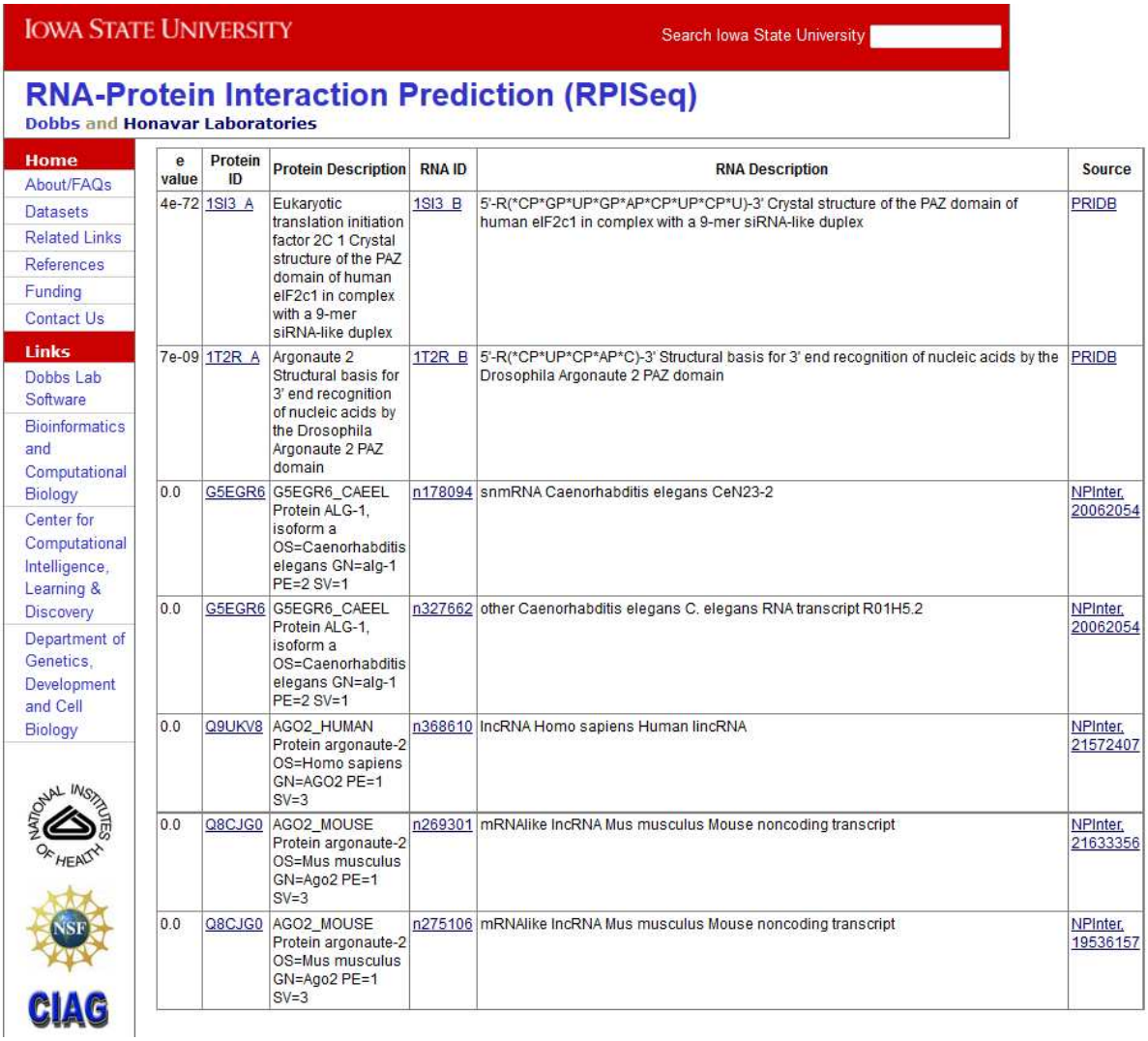


Figure 4.1 A. Sample output of RPISeq webserver with a single protein and a single RNA. B. Sample output of a batch submission predictions with a single RNA and multiple proteins.

A current limitation of the RPISeq v 2.0 server is that input is limited to 100 protein sequences or 100 RNA sequences during batch submission, with a maximum file size of ~1.5 MB. If users are interested in running predictions on a larger scale, they can request a free local implementation of RPISeq.

RPIntDB

Another important enhancement implemented in the RPISeq v 2.0 server is seamless integration with a newly developed database, the RNA-Protein Interaction DataBase (RPIntDB) (http://pridb.gdcb.iastate.edu/RPISeq/RPIntDB.html). This feature allows users to query a protein sequence against a large collection of experimentally validated RNA-protein interactions. RPIntDB contains a total of 44,586 RNA-protein interactions, comprising 11,928 unique RNAs and 2190 unique proteins. It includes interactions from structurally characterized RNA-protein complexes in the PRIDB (Lewis *et al.*, 2011), as well as individual and high throughput experiments extracted from the NPInter database (Wu *et al.*, 2006). For RPIntDB queries, RPISeq accepts a single protein sequence as input. The protein sequence is used as the query in a BLAST search (Altschul *et al.*, 1990) against all protein sequences in RPIntDB. The resulting protein hits, together with their known RNA interaction partners, are returned to the user. When querying against the database, the user can adjust the *e*-value for the BLAST search to either improve the specificity of BLAST hits (i.e., reduce false positives) or enhance the search sensitivity to improve detection of remote homologs. Sample output from an RPIntDB search is shown in Figure 4.2. For each hit in the output, links are provided to additional information about the protein, RNA and source of each interaction.



e value	Protein ID	Protein Description	RNA ID	RNA Description	Source
4e-72	1SI3_A	Eukaryotic translation initiation factor 2C 1 Crystal structure of the PAZ domain of human eIF2c1 in complex with a 9-mer siRNA-like duplex	1SI3_B	5'-R(*CP*GP*UP*GP*AP*CP*UP*CP*U)-3' Crystal structure of the PAZ domain of human eIF2c1 in complex with a 9-mer siRNA-like duplex	PRIDB
7e-09	1T2R_A	Argonaute 2 Structural basis for 3' end recognition of nucleic acids by the Drosophila Argonaute 2 PAZ domain	1T2R_B	5'-R(*CP*UP*CP*AP*C)-3' Structural basis for 3' end recognition of nucleic acids by the Drosophila Argonaute 2 PAZ domain	PRIDB
0.0	G5EGR6	G5EGR6_CAEEL Protein ALG-1, isoform a OS=Caenorhabditis elegans GN=alg-1 PE=2 SV=1	n178094	snmRNA Caenorhabditis elegans CeN23-2	NPInter, 20062054
0.0	G5EGR6	G5EGR6_CAEEL Protein ALG-1, isoform a OS=Caenorhabditis elegans GN=alg-1 PE=2 SV=1	n327662	other Caenorhabditis elegans C. elegans RNA transcript R01H5.2	NPInter, 20062054
0.0	Q9UKV8	AGO2_HUMAN Protein argonaute-2 OS=Homo sapiens GN=AGO2 PE=1 SV=3	n368610	lncRNA Homo sapiens Human lincRNA	NPInter, 21572407
0.0	Q8CJG0	AGO2_MOUSE Protein argonaute-2 OS=Mus musculus GN=Ago2 PE=1 SV=3	n269301	mRNAlike lncRNA Mus musculus Mouse noncoding transcript	NPInter, 21633356
0.0	Q8CJG0	AGO2_MOUSE Protein argonaute-2 OS=Mus musculus GN=Ago2 PE=1 SV=3	n275106	mRNAlike lncRNA Mus musculus Mouse noncoding transcript	NPInter, 19536157

Figure 4.2 Sample RPIntDB results.

The RPISeq 2.0 webserver provides a tutorial that explains how to use RPISeq and includes examples of results obtained for typical queries. Sample protein and RNA sequences are provided on each input page. The datasets used in RPISeq are freely available for download. In addition, all interactions in RPIntDB can be downloaded as a tab-delimited file.

RPISeq 2.0 runs on the Apache 2.2 webserver, using MySQL 14.14 as a database backend for RPIntDB and PHP 5 for user interface functions. Input processing and prediction algorithms are implemented using Perl 5 scripts and Weka 3.7.1 (Hall et al., 2009). In

addition to the RPISeq 2.0 webserver (http://pridb.gdcb.iastate.edu/RPISeq), a local implementation of RPISeq is freely available upon request.

References

Agostini F, Zanzoni A, Klus P, Marchese D, Cirillo D, Tartaglia GG: catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics* 2013.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403-410.

Bellucci M, Agostini F, Masin M, Tartaglia GG: Predicting protein associations with long noncoding RNAs. *Nature Methods* 2011, 8:444-445.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: The WEKA data mining software: An update. *SIGKDD Explorations* 2009, 11:10-18.

Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D: PRIDB: a Protein-RNA Interface Database. *Nucleic Acids Res* 2011, 39:D277-82.

Lu Q, Ren S, Lu M, Zhang Y, Zhu D, Zhang X, Li T: Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics 2013, 14:651.

Muppirala UK, Honavar V, Dobbs D: Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics* 2011, 12:489.

Muppirala UK, Lewis BA, Dobbs D: Computational tools for investigating RNA-protein interaction partners. *J Comput Sci Syst Biol* 2013, 6:182-187.

Pancaldi V, Bähler J: In silico characterization and prediction of global protein-mRNA interactions in yeast. *Nucleic Acids Res* 2011, 1-11.

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H: Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci* USA 2007, 104:4337-41.

Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, El-Manzalawy Y, Dobbs D, Honavar V: Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics* 2012, 13:89.

Wang Y, Chen X, Liu ZP, Huang Q, Wang Y, Xu D, Zhang XS, Chen R, Chen L: De novo prediction of RNA-protein interactions from sequence information. *Mol Biosyst* 2013, 9: 133-142.

Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, Zhang Z, Skogerbø G, Chen L, Lu H, Zhao Y, Chen R: NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res* 2006, 34:D150-2.

CHAPTER 5. A MOTIF-BASED METHOD FOR PREDICTING INTERFACIAL RESIDUES IN BOTH THE RNA AND PROTEIN COMPONENTS OF PROTEIN-RNA COMPLEXES

Abstract

Efforts to predict interfacial residues in protein-RNA complexes have largely focused on predicting RNA binding residues in proteins. Predicting residues on the RNA side of the interface, i.e., predicting protein binding residues in RNA sequences, is a problem that has received little attention to date. Although the value of sequence motifs for classifying and annotating protein sequences is well established, sequence motifs have not been widely applied to predicting interfacial residues in macromolecular complexes. Here, we propose a novel sequence motif-based method for "partner-specific" interfacial residue prediction. Given a protein-RNA pair, the goal is to simultaneously predict RNA binding residues in the protein sequence and protein binding residues in the RNA sequence. In 5-fold cross validation experiments, our method, PS-RPIMotif, achieved a specificity of 92% and a sensitivity of 61%, with correlation coefficient (CC) of 0.58 in predicting RNA-binding sites in proteins. The method achieved 69% specificity and 75% sensitivity, but with a low CC of 0.13 in predicting protein binding sites in RNAs. Similar results were obtained when PS-RPIMotif was tested on an independent "blind" dataset of 327 protein-RNA interactions, suggesting the method should be widely applicable and valuable for the identifying potential

interfacial residues in protein-RNA complexes for which structural information is not available.

Introduction

Despite the important roles of protein-RNA interactions in many cellular activities, including transcription, translation, viral replication and pathogen resistance (Hogan *et al.*, 2008, Licatalosi *et al.*, 2010, Kim *et al.*, 2009, Sola *et al.*, 2011), the determinants of protein-RNA recognition are not yet fully understood. The Protein Data Bank (PDB) (Berman *et al.*, 2000) is a valuable resource for studying protein-RNA complexes, but the number of protein-RNA complex structures available in the PDB is less than 1% of the total structures. Even so, these data have been successfully exploited to develop several computational methods for predicting interfacial residues in protein-RNA complexes (Jeong *et al.*, 2004, Terribilini *et al.*, 2006, Maetschke and Yuan, 2009, reviewed in Puton *et al.*, 2012, Walia *et al.*, 2013) and, recently, a few methods for predicting interaction partners in protein-RNA interaction networks (Muppirala *et al.*, 2011, Bellucci *et al.*, 2011, reviewed in Muppirala *et al.*, 2013).

Methods for predicting RNA-binding residues in proteins fall into two major classes: i) methods that use only sequence information and ii) methods that take advantage of structural information, when available (Puton *et al., 2012,* Walia *et al.,* 2013). None of the published methods, with one exception, Choi and Han (2010), take into account information regarding the RNA partner, i.e., they are non-partner-specific predictors of interfacial residues.

Computational prediction of protein-binding RNA nucleotides is an even harder problem (Choi and Han, 2013). Due in part to the limited 4-nucleotide alphabet of RNA (and

its consequently low per-character information content), studies that have attempted to draw more general conclusions about protein-RNA interactions have focused on the protein side of the interface. Many analyses of RNA sequence have focused on specific features involved in cellular pathways, such as ribosome binding sites (Chang *et al.*, 2013). While some small structural elements in RNAs have been elucidated (Fritsch and Westhof, 2010, Petrov *et al*., 2013), examination of these motifs has focused primarily on their roles in mediating RNA-RNA contacts in the context of a larger RNA structure, with few studies considering the interaction between RNA structural motifs and proteins (Ciriello *et al.*, 2010).

Here, we perform a large scale analysis of contiguous sequence motifs present in the interfaces of protein-RNA complexes and develop a new "partner-specific" motif-based method to simultaneously predict RNA binding residues in the protein component and protein binding ribonucleotides in the RNA component of a given protein-RNA pair.

Methods

Generating interfacial sequence motifs

To generate interfacial sequence motifs with which to scan target sequences, a dataset of all protein-RNA complex structures deposited in the Protein Data Bank (PDB) as of September 2012 was analyzed to find short regions, contiguous in primary sequence, and composed entirely of interacting (as defined using an 8Å distance cutoff) residues in either the protein or RNA chains. The sequences of these interfacial segments (without any information about their interacting partner residues) were extracted as '*n*-mer motifs', where *n* can vary between 3 and 8. No requirement was made for motifs to be bounded by non-

interacting residues; therefore overlapping motifs were included. Note that a 5-mer motif necessarily contains two 4-mer motifs and three 3-mer motifs.

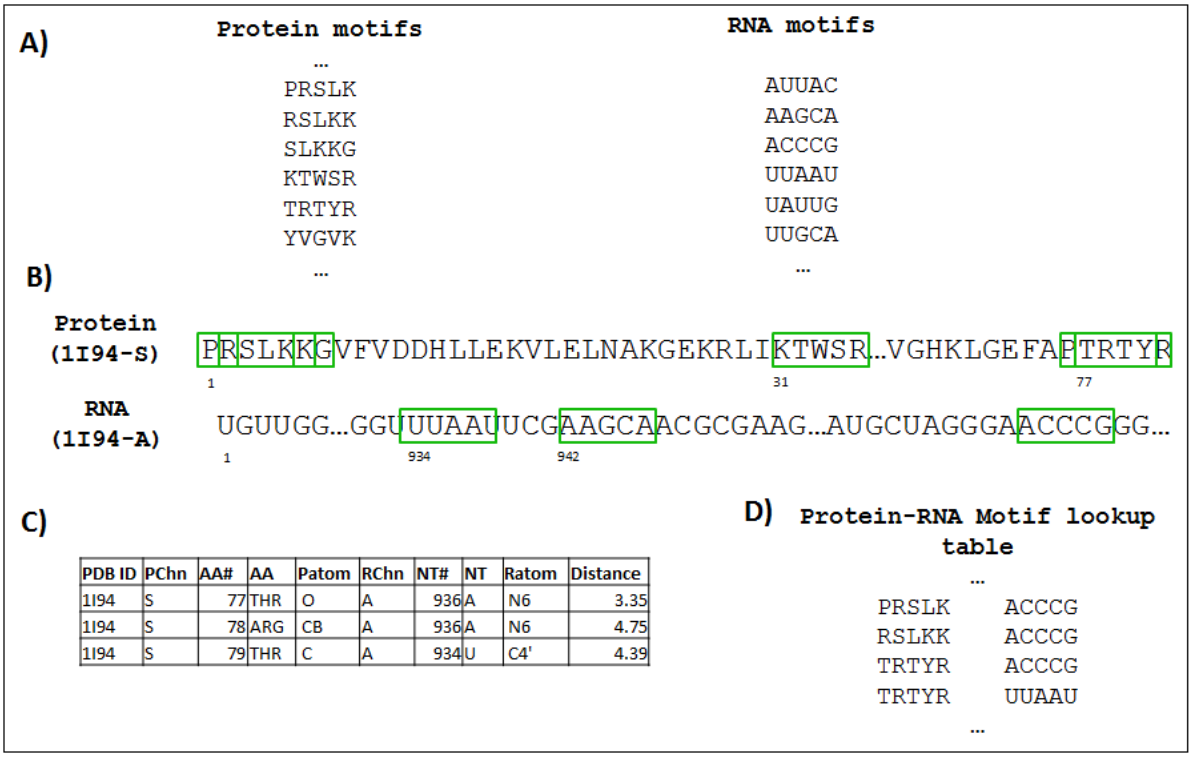
Datasets for interface prediction

To generate a dataset for evaluating the utility of motifs for interface prediction, interacting protein and RNA chains were extracted from ribosomal complexes with at least 3.5Å resolution. Proteins of length less than 25 amino acids and RNAs of length less than 100 ribonucleotides were not included. The interaction information (i.e., interfacial residues) for these chains was downloaded from PRIDB (Lewis *et al.*, 2011). For this dataset, residues in protein and RNA chains were defined as interacting if any atom in one chain lies within a 5Å distance cutoff from any atom in the other chain. Redundant protein sequences (i.e., with $\geq$ 30% sequence identity across the entire length) interacting with similar RNA sequences (i.e., with $\geq$ 30% sequence identity across the entire length) were discarded and vice-versa for redundant RNA sequences. This resulted in a total of 1,637 interacting protein-RNA pairs. 327 pairs were kept aside for independent evaluation and 5-fold cross-validation was performed on the remaining 1,310 pairs.

Generating a protein-RNA interface motif lookup table

The protein-RNA interface motif lookup table consists of pairs of protein and RNA interfacial sequence motifs that are known to contact one another (i.e., to have at least one amino acid-ribonucleotide interaction) in a characterized protein-RNA complex. Entries in the lookup table were obtained as follows: First, the protein sequences in all known protein-RNA pairs were scanned for interfacial sequence motifs (identified as described above) using a sliding window approach. Similarly, RNA sequences were scanned for interfacial sequence

motifs. Second, every pair of protein-RNA sequences in the dataset of known protein-RNA interactions was scanned to identify cases in which there exists at least one physical interaction between the amino acids and ribonucleotides of a corresponding pair of sequence



PDB ID	PChn	AA#	AA	Patom	RChn	NT#	NT	Ratom	Distance
1I94	S	77	THR	O	A	936	A	N6	3.35
1I94	S	78	ARG	CB	A	936	A	N6	4.75
1I94	S	79	THR	C	A	934	U	C4'	4.39

motifs. If an interaction is observed, that particular protein-RNA sequence motif pair is added to the lookup table. This method is further explained in Figure. 5.1.

Figure 5.1 Generating the protein-RNA motif lookup table. A) A subset of the protein and RNA interfacial motifs used to scan target sequences are shown. B) The protein and RNA sequence of each protein-RNA pair in the training dataset are scanned with these interfacial motifs. For the purpose of illustration, only a portion of the example sequences and a subset of the interfacial motifs (indicated in green boxes) are shown. C) PRIDB is used to identify interacting residues within an distance threshold of 5A. Only a subset of interactions identified in this example are shown. D) Only protein and RNA motif pairs which contain at least one such interaction between them are added to the protein-RNA motif

lookup table. Of the eighteen possible protein-RNA motif pairs illustrated in this example, only four satisfy this criterion and are added to the lookup table.

Motif-based prediction of interfacial residues in both RNA and protein

After generating the protein-RNA interface motif lookup table, prediction of interfacial residues in a query protein-RNA pair was done in a single step. The protein and RNA sequences were scanned simultaneously for the presence of all motif pairs in the lookup table. If any motif pair is present, those amino acids and ribonucleotides are marked as "interfacial" in the given sequences. The remaining residues and ribonucleotides are marked as non-interfacial residues. For example, using the lookup table in Figure 5.1, if 'TRTYR' is found in the query protein and 'UUAAU' is found in the query RNA, the corresponding amino acids and ribonucleotides are predicted as interfacial residues.

Performance evaluation

We used the following measures to evaluate the performance of motif-based prediction of interfacial residues on both proteins and RNAs. TP (true positives) refers to the number of interface residues correctly identified as such by the method. FP (false positives) refers to the number of non-interface residues misclassified as interface residues. FN (false negatives) refers to the number of interface residues misclassified as non-interface residues. TN (true negatives) refers to the number of non-interface residues correctly identified as such by the method.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Results

Distribution of interfacial amino acid motifs in proteins from known protein-RNA complexes

Although there are 20^5 different potential combinations of amino acid 5-mers, only 0.3% (11,269) of the theoretically possible motifs were observed in interfaces extracted from known protein-RNA complexes (1,408 complexes, comprising 17,385 protein chains) in the PRIDB database (Lewis *et al.*, 2011). In this comprehensive dataset of interfaces, certain motifs, such as 'AKTDS', 'LPVRG' and 'DPHPG', are highly over-represented relative to other motifs (data not shown). To examine the frequency distribution of these motifs, we extracted motifs from a non-redundant subset of the above dataset, which was generated using Blastclust with a 30% sequence identity cutoff. The motif frequency distribution plot obtained for the non-redundant dataset is shown in Figure 5.2. Several peaks corresponding to interfacial motifs that occur at a high frequency are observed (e.g., AKTDS, LPVRG, QYAKT). Approximately 50% of the protein 5-mer motifs are observed only once (as a contiguous stretch of interfacial residues) in the interfaces.

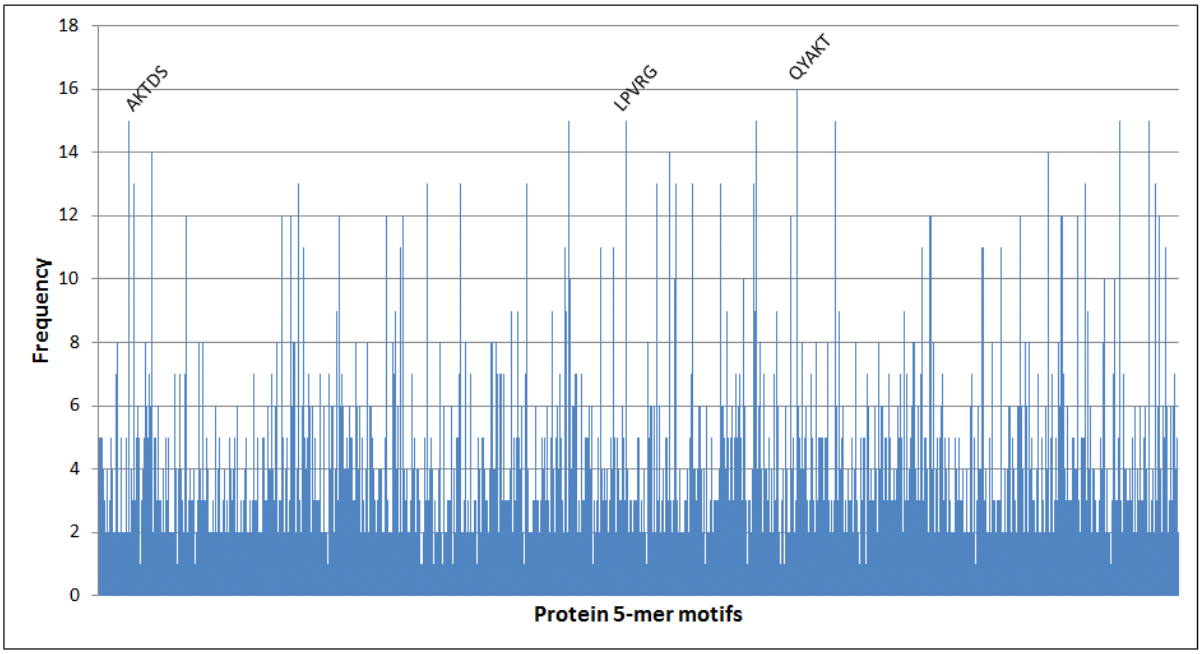


Figure 5.2 Frequency distribution of protein 5-mer motifs in non-redundant protein sequences in PRIDB. Some of the peaks corresponding to highly represented motifs such as 'AKTDS' and 'LPVRG' are labeled.

Motif-based partner-specific prediction of interfacial residues

To evaluate whether an interface motif lookup table can be used to predict interfacial residues in specific protein-RNA pairs, we first performed preliminary experiments in which we tested the effect of varying the length of protein motifs from 4 to 6 amino acids, and the length of RNA motifs from 4 to 8 ribonucleotides (see *Methods*). As expected, using shorter motifs resulted in a larger number of false positive predictions, whereas using longer motifs resulted in larger number of false negative predictions. Based on these results, we determined that a protein motif of length 5 should provide a good balance between prediction specificity and sensitivity.

To predict RNA binding residues in the protein component of a given protein-RNA pair, we used a protein motif size of length 5 and varied the RNA motif lengths from 4 to 6.

Table 5.1 summarizes the average prediction results obtained using a 5-fold cross validation approach, in which 80% of the data was used to generate the protein-RNA motif lookup table and predictions were made on the remaining 20% of the data. There is little difference in the specificity or correlation coefficient (CC) using RNA motifs of length 4 and 5. Although using an RNA motif of length 6 resulted in higher specificity (0.94), it resulted in lower sensitivity and CC compared with using RNA 4- and 5-mers. Using an RNA 4-mer resulted in higher sensitivity (0.65) compared with using 5- and 6-mers.

Table 5.1 RNA-binding residue prediction performance using 5-fold cross validation on a non-redundant dataset of 1,310 protein-RNA pairs

Protein Motif length	RNA motif length	Specificity	Sensitivity	CC
5	4	0.90	0.65	0.58
5	5	0.92	0.61	0.58
5	6	0.94	0.54	0.54

To predict which ribonucleotides in the RNA component of a given protein-RNA pair participate in protein binding, we again used a protein motif size of length 5 and varied the RNA motif lengths from 4 to 6. Table 5.2 summarizes the prediction results obtained in 5-fold cross-validation experiments. Again, as the RNA motif size is increased, the specificity increased, but with the expected decrease in sensitivity. A high specificity of 0.91 is obtained using an RNA motif length of 6, but the corresponding CC is much lower than that obtained for RNA binding site prediction (Table 5.1).

Table 5.2 Protein-binding residue prediction performance using 5-fold cross validation on a non-redundant dataset of 1,310 protein-RNA pairs

Protein Motif length	RNA motif length	Specificity	Sensitivity	CC
5	4	0.35	0.89	0.07
5	5	0.69	0.75	0.13
5	6	0.91	0.55	0.21

Prediction on an independent test set

To more rigorously test the performance of the method, we evaluated it on an independent dataset of 327 protein-RNA pairs (See *Methods*). As summarized in Table 5.3, using protein and RNA motifs of length 5, we obtained 92% specificity and 64% sensitivity in predicting RNA binding residues. In predicting protein binding ribonucleotides, the specificity was 67% and sensitivity was 79%. Thus, performance on the independent test set was comparable to that obtained in cross-validation experiments. This suggests that our proposed “partner-specific” method for predicting RNA-protein interfaces using sequence motifs, which we call PS-RPIMotif, should be generally applicable.

Table 5.3 Prediction performance on an independent test set of 327 protein-RNA pairs **using protein and RNA motifs of length 5.**

Prediction	Specificity	Sensitivity	CC
RNA binding amino acids in proteins	0.92	0.64	0.59
Protein binding nucleotides in RNA	0.67	0.79	0.13

Comparison with other interface prediction methods

Only one other published study has addressed the prediction of binding sites in proteins and RNAs simultaneously. The catRAPID method proposed by Bellucci *et al.* (2011) divides the protein and RNA sequences into a number of fragments and calculates interaction propensities between each pair of protein-RNA fragments. Binding site prediction on a per residue basis was not reported. Because neither the details of the method nor the performance evaluation results were reported, we cannot compare our PS-RPIMotif method with catRAPID.

The only other published method for predicting protein binding sites in RNAs was reported by Choi and Han (2013). Unfortunately, we have not been able to make direct performance comparisons with their method because neither their test dataset nor a working webserver is available. In an earlier report, Choi and Han also proposed a partner-specific RNA binding site prediction method, in which the RNA sequence is encoded as the sum of the normalized positions of each nucleotide (A, C, G and U) in the sequence (Choi and Han, 2010). When we examined the dataset used in that study, we noticed that all except one RNA sequence was less than 100 nucleotides in length, and approximately half of the dataset

consists of very short RNAs (< 15 nts). Because the minimum length of the RNA used in our training dataset is 100 nt, and, as discussed in the next section, our method is not suitable for small RNAs, we did not compare PS-RPIMotif with Choi and Han's method. There is no webserver implementing the Choi and Han method and we did not attempt to re-implement it in order to provide a direct comparison with our method. Choi and Han reported prediction performance of 91% specificity, 60.7% sensitivity with a CC of 0.24 on a dataset of 267 interacting protein-RNA pairs (Choi and Han, 2010).

We were able to compare the performance of our *partner-specific* PS-RPIMotif method with existing *non-partner* specific sequence-based methods for predicting RNA binding residues in proteins. Walia *et al.* (2013) performed a systematic comparison of existing methods for predicting RNA binding residues and showed that PSSM-based methods had the best performance among published sequence-based approaches. Thus, we directly compared the performance of PS-RPIMotif with RNABindRPlus (Walia *et al.*, in preparation), which combines homology-based predictions with predictions from an optimized SVM classifier that uses a PSSM-based approach. Because homology-based methods exploit existing structures and interfaces, and our independent test set was extracted from the PDB, we expected the homology-based method to perform very well. Homology-based methods fail, however, when the query sequence has no homologs in the PDB. We also compared our method with the SVM component of RNABindRPlus and the results are also shown in Table 5.4. PS-RPIMotif has better performance in terms of specificity (0.92), but lower sensitivity (0.64) compared to RNABindRPlus. RNABindRPlus had the highest CC (0.71); the CCs for the other two methods were similar (0.59 vs 0.61). A larger difference is

seen in the precision (or positive prediction rate) of the two methods: PS-RPIMotif has higher precision (0.80) than RNABindRPlus (0.76) on this dataset.

Table 5.4 Performance comparison of PS-RPIMotif and RNABindRPlus in the prediction of RNA binding sites.

Method	Specificity	Sensitivity	Precision	CC
PS-RPIMotif	0.92	0.64	0.80	0.59
RNABindRPlus	0.85	0.88	0.76	0.71
RNABindRPlus (SVM-only)*	0.74	0.90	0.65	0.61

* The RNABindRPlus (SVM-only) did not return predictions for protein chains 1W2B_G, 3D5B_J and 1JJ2_G; it failed to generate PSSMs for these sequences.

Discussion

Our results indicate that specific subsets of short contiguous interfacial motifs are over-represented relative to other interfacial motifs within the sequences of both protein and RNA components of protein-RNA complexes. A large number of interfacial amino acid motifs occur only once in the dataset analyzed here. This may be a consequence of the criteria for generating the short RNA binding motifs in this study: all residues in an interfacial motif must be contiguous in sequence and must interact with at least one atom in a ribonucleotide within a 5 Å distance cutoff. It is striking that a simple lookup table of motif pairs, identified in a training set of protein-RNA complexes, can be used to accurately predict interfacial residues in an independent set of complexes. Although we have not yet directly calculated the interface propensities of these motifs (i.e., the over-representation of these motifs in interfacial versus non-interfacial regions of the protein and RNA sequences), it may

be possible to improve prediction of interfacial residues by focusing on motifs with high interface propensity.

The interface prediction results reported here demonstrate that an RNA motif of length 5, while not informative on its own, can be highly informative when used in combination with a protein motif of length 5. From the non-redundant dataset of RNA-protein complexes used in this study, we generated a lookup table of 55,154 protein-RNA motif pairs, comprising 3,275 unique protein motifs and 835 unique RNA motifs. Using a non-redundant dataset is the appropriate way to evaluate and compare interface prediction methods, but doing so is expected to exclude some informative motif combinations. Thus, we created a motif lookup table *without* discarding redundant motifs. As expected, many additional protein-RNA motif pairs were identified: a total of 88,994 protein-RNA motif pairs, comprising 4,035 protein motifs and 893 RNA motifs.

Our results indicate that binding partner information, which has been largely ignored for predicting interfacial residues in protein-RNA complexes, can be valuable for making "partner-specific" interface predictions. Figure 5.3 illustrates this with an example. In the *E. coli* ribosome, 16S rRNA in the small subunit interacts with various protein components of the 30S subunit, using different binding sites. Interaction of S4 and S11 proteins with a segment of 16S ribosomal RNA (PDB 4GAS) is shown in the inset of Figure 5.3. In this structure, the majority of 16S rRNA nucleotides that bind the S4 protein are located in the region 400 – 440. In contrast, region 670 – 720 of 16S RNA contains most of S11 protein binding residues. In 16S RNA, different interface predictions are obtained for the S4 and S11 proteins. As shown in Figure 5.3, many interfacial residues are correctly predicted in the S4

binding region, while on the same regions, where S11 protein does not bind, only a few residues are incorrectly predicted as interfacial (i.e., are false positive predictions).

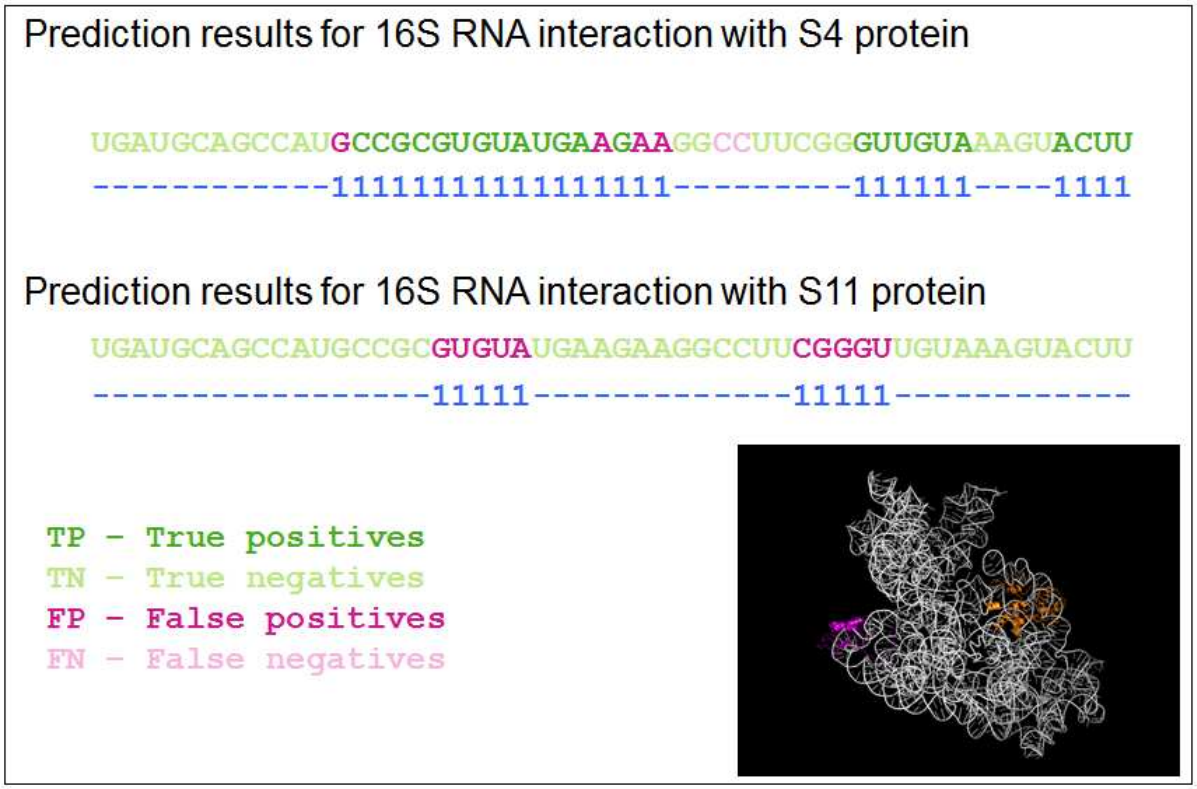


Figure 5.3 Example of a partner-specific interface prediction. Different interfacial residues (protein-binding residues) are predicted for the same RNA sequence (Residues 386-437 of 16S RNA (PDB ID: 4GAS)), when it is paired with two different protein partners (S4 protein and S11 protein). The predictions are indicated by '1' and '-'. The inset picture shows the structure of 16S ribosomal RNA bound to proteins S4 and S11. Residues 386-437 are part of the S4 binding region.

PS-Hom-PPI is a partner-specific homology-based method for predicting protein-protein interaction sites; it predicts interfacial residues in both partners of a query protein-protein complex by identifying homologous protein pairs for which a complex structure is available (Xue *et al.*, 2011). The method performs very well if a protein complex homologous to the query can be identified in the PDB. PS-RPIMotif takes into account the

partner information, as does PS-Hom-PPI, but differs in that it uses only small sequence motifs to scan the inputs.

A limitation of the current PS-RPIMotif method is that it cannot predict interfacial regions of lengths shorter than 5 residues because the minimum length of motifs used for scanning the sequences is 5. In particular, the current implementation cannot accurately predict interface residues in very short RNAs. Short RNAs (which often correspond to interface-containing fragments of much longer RNAs present in native complexes) are common in structurally-characterized protein-RNA complexes in the PDB. Thus, the likelihood that every ribonucleotide in such an RNA is an interfacial residue is very high compared to the situation for longer RNAs, in which a only a small fraction of the ribonucleotides directly contact the bound protein(s). Because of this short RNA bias in the PDB, we excluded RNAs less than 100 nts in length for generating our motifs (see *Methods*). In our experiments, PS-RPIMotif performed well on RNAs greater than 100 nts in length, but poorly when tested on RNAs shorter than 100 nts (data not shown). Thus, PS-RPIMotif can be used to predict protein-binding sites in mRNAs, rRNAs, long non-coding RNAs and many short ncRNAs, but predictions on RNAs less than 100 nts are likely to be unreliable.

In future work, we plan to evaluate effect of incorporating predicted RNA secondary structure in the RNA sequence representation, which may lead to better performance in predicting protein binding residues in RNA. Currently, we are evaluating whether our motif-based approach can be applied to the partner prediction problem (i.e., predicting whether or not a given protein-RNA pair will interact). A webserver for PS-RPIMotif is under construction and will be available soon.

Conclusion

We have developed a new method for predicting partner-specific interfacial residues in protein-RNA complexes using short sequence motifs. PS-RPIMotif can simultaneously predict interfacial residues in both the protein and RNA components of a complex. An RNA motif of length 5, in combination with a protein motif of length 5, can be used to predict interfacial residues with high specificity (0.92 for RNA binding residues in proteins; 0.67 for protein binding residues in RNA), indicating that PS-RPIMotif can be a valuable tool for experimentalists who wish to target interfaces in specific protein-RNA complexes or to perturb specific interactions in protein-RNA interaction networks.

References

Bellucci M, Agostini F, Masin M, Tartaglia GG: Predicting protein associations with long noncoding RNAs. Nat Methods 2011, 8: 444-445.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. Nucleic Acids Res 2000, 28:235-42.

Chang TH, Huang HY, Hsu JB, Weng SL, Horng JT, Huang HD: An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. *BMC Bioinformatics* 2013, 14(Suppl 2): S4.

Choi S, Han K: Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. Computers in Biology and Medicine 2013, 43(11): 1687-97.

Choi S, Han K: Prediction of RNA-binding amino acids from protein and RNA sequences. BMC Bioinformatics 2011, 12: S7.

Ciriello G, Gallina C, Guerra C: Analysis of interactions between ribosomal proteins and RNA structural motifs. *BMC Bioinformatics* 2010, 11(Suppl 1): S41.

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR: RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 2011, 39: D301-308.

Fritsch V, Westhof E: The architectural motifs of folded RNAs. The Chemical Biology of Nucleic Acids. (2010).

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO: Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol 2008, 6:e255.

Jeong E, Chung I, Miyano S: A neural network method for identification of RNA-interacting residues in protein. Genome Inform 2004, 15:105-116.

Kim MY, Hur J, Jeong S: Emerging roles of RNA and RNA-binding protein network in cancer cells. BMB Rep 2009, 42:125-130.

Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D: PRIDB: a Protein-RNA Interface Database. Nucleic Acids Res 2011, 39:D277-82.

Licatalosi DD, Darnell RB: RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 2010, 11:75-87.

Maetschke S, Yuan Z: Exploiting structural and topological information to improve prediction of RNA-protein binding sites. BMC Bioinf 2009, 10:341

Muppirala UK, Honavar V, Dobbs D: Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics* 2011, 12:489.

Muppirala UK, Lewis BA, Dobbs D: Computational tools for investigating RNA-protein interaction partners. J Comput Sci Syst Biol 2013, 6:182-7.

Petrov AI, Zirbel CL, Leontis NB: Automated classification of RNA 3D motifs and the RNA 3D motif atlas. RNA 2013, 19:1327-1340.

Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM: Computational methods for the prediction of protein-RNA interactions. J Struct Biol 2012, 179(3):261-8.

Sola I, Mateos-Gomez PA, Almazan F, Zuñiga S, Enjuanes L: RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. RNA Biol 2011, 8:237-248.

Terribilini M, Lee J, Yan C, Jernigan R, Honavar V, Dobbs D: Prediction of RNA-binding sites in proteins from amino acid sequence. RNA 2006, 16(12):1450-1462.

Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, El-Manzalawy Y, Dobbs D, Honavar V: Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinformatics 2012, 13:89.

Walia RR, Xue LC, Wilikins K, El-Manzalawy Y, Dobbs D, Honavar V: RNABindRPlus: A Sequence Homology-Based Approach to Predicting RNA-binding Sites in Proteins and its Combination with Machine Learning Methods. (In preparation).

Xue LC, Dobbs D, Honavar V: HomPPI: a class of sequence homology based protein-protein interface prediction methods. BMC Bioinformatics 2011, 12:244.

CHAPTER 6. GENERAL CONCLUSIONS

Protein-RNA interactions are responsible for regulating a wide variety of cellular processes. Characterization of these interactions, including identification of RNA-protein interaction partners and interfacial residues in proteins and RNAs, is essential for understanding how these processes are regulated. In this dissertation, we have presented: i) a new method for predicting protein-RNA interaction partners; ii) a webserver for predicting partners; iii) a comprehensive database of known protein-RNA interactions; and iv) a new "partner-specific" method for predicting interfacial residues on proteins and RNAs simultaneously.

Contributions

Classifiers that predict RNA-protein interaction partners

We developed a novel sequence-based machine learning method to predict whether a given protein and RNA interact (Muppirala *et al.*, 2011). We demonstrated that, at least for a large dataset of protein-RNA complexes extracted from the PDB, the protein and RNA sequences alone (i.e., without taking advantage of any available structural or functional information) contain enough signal to allow reliable prediction of interaction partners. Our method was also shown to perform well on an independent dataset of RNA-protein interactions extracted from NPInter (Wu *et al.*, 2006), and to accurately predict ncRNA-protein interaction networks. Our approach can be used to predict either putative RNA partners for a target protein or putative protein partners for a target RNA. One of the limitations of this method is a high number of false positives when tested on non-RNA

binding proteins. This can be overcome by using real negative examples to train the classifiers.

A webserver for predicting binding partners of proteins or RNAs

We developed RPISeq, a server for predicting the interaction probability of a given protein-RNA pair (http://pridb.gdcb.iastate.edu/RPISeq/). RPISeq allows users to submit multiple protein or RNA sequences and to make predictions on a large scale. RPISeq has been accessed thousands of times from 25 countries. One recent study used the RPISeq webserver to identify linc-UBC1 RNA as a potential interaction partner of the PRC2 (Polycomb Repressive Complex 2) protein; this prediction was experimentally validated using RNA immunoprecipitation (He *et al.*, 2013). At this time, since the webserver restricts the length of input proteins and RNAs, it is not possible to run large scale predictions on the entire proteome or transcriptome of an organism. However, users can request an offline version of the program is available upon request.

A comprehensive database of RNA-protein interactions

We developed RPIntDB, a comprehensive database of RNA-protein interactions, which is integrated with the RPISeq webserver. RPIntDB is a collection of interactions from existing literature and databases, such as PRIDB (Lewis *et al.*, 2011) and NPInter (Wu *et al*, 2006), Currently, RPIntDB contains 44,586 interactions comprising 2190 unique proteins and 11,928 unique RNAs. Queries of RPIntDB can be used to complement or corroborate RPISeq predictions: users can identify potential RNA partners for a protein of interest based on a BLAST search against protein sequences in RPIntDB. The search returns homologous protein sequences for which interacting RNA partners are known. Taken together, RPISeq

and RPIntDB are valuable resources for those interested in studying protein-RNA interaction partners.

A motif-based method for "partner-specific" interface residue prediction

With the goal of identifying sequence motifs potentially predictive of protein-RNA interfaces, we performed an analysis of contiguous interfacial amino acids and ribonucleotides in protein-RNA complexes in the PRIDB. We showed that certain protein 5-mers occur more frequently than others in interfaces. Based on this result, we a developed a novel sequence motif-based method that simultaneously predicts interfacial residues in both the protein and RNA partners of a complex. We demonstrated that protein 5-mer motifs, in combination with RNA 5-mer motifs, can be used to predict "partner-specific" interfacial residues, and that using available binding partner information leads to higher precision in the prediction of RNA-binding amino acids in proteins.

Future Studies

Predicting protein-RNA interfaces and interaction partners are challenging problems. Especially, predicting protein binding residues in RNA is a very hard problem that has received very little attention to date, and the predictions we have obtained so far are not optimal. There are several avenues to pursue to build on the work presented in this dissertation to improve both the prediction of protein-RNA interfaces and the prediction of interaction partners.

Improving the prediction of interaction partners by RPISeq: One limitation mentioned above is the lack of validated "negative" examples for training classifiers that

predict interaction partners. In training the RPISeq classifiers, positive examples were derived from proteins and RNAs found in structurally characterized complexes (i.e., from the PDB). Negative examples were generated by randomly pairing the same set of proteins and RNAs (and removing any pairs that were present in the positive set). Making use of real, experimentally-validated negative examples identified in high throughput RNA-binding experiments such as RIP-Chip (Keene *et al.*, 2009) would be expected to improve prediction accuracy.

For protein-protein interactions, the Negatome (Smialowski *et al.*, 2010), is a database of protein and protein domain pairs that are unlikely to be engaged in direct physical interactions. For protein-RNA interactions, no such database exists. It would be useful to have a repository of non-interacting protein-RNA pairs. Information on non-interacting pairs can also be obtained from some high-throughput experiments (Ray *et al.*, 2009). In future, we plan to provide a user-friendly interface through which researchers can submit their interaction data (both positive and negative examples) for incorporation in the RPIntDB database. Submitted information will be curated and added into the database.

Further development of RPIntDB: The current implementation of RPIntDB allows users to input a single protein sequence to obtain homologous proteins and their corresponding RNA partners. In the future, we plan to provide search functionalities that will enable users to search for specific RNA sequences as well. We also plan to provide options to filter the search results based on the source of interactions.

Increase distance cutoff for predicting interfaces: For prediction of interfacial residues, we have obtained the sequence motifs using a distance cutoff of 5Å. While this cutoff is sufficient to capture many types of short range interactions, it may miss important

longer range interactions such as electrostatic interactions (typically ~8 Å). By increasing the distance cutoff, we may be able to capture more interaction signals.

Develop a webserver for "partner-specific" prediction of interfacial residues: Application of any prediction method is limited if there is no available webserver or an easy way to reproduce the method. We are developing a new webserver that implements the partner-specific interfacial residue prediction method. Users will input a pair of potentially interacting protein and RNA sequences. Output will provide predicted interfacial residues labeled as '+' and non-interfacial residues labeled as'-'. We will provide provision for batch submission of multiple protein-RNA pairs. We will also allow download of the protein-RNA motif lookup table.

Develop a multi-stage classifier for predicting interaction partners and interfaces: We have developed two methods for i) predicting protein-RNA interaction partners and ii) interfacial residues. Future work should include combining these methods to generate a multi-stage classier. First, a given protein can be tested for its RNA-binding propensity. If it is predicted to be an RNA binding protein, then its interaction with a given RNA(s) can be tested. If there is a high probability of interaction with a specific RNA sequence, the interfacial residues can be predicted using the motif-based method (or other methods developed in our group for predicting interfacial residues in RNA binding proteins).

References

He W, Cai Q, Sun F, Zhong G, Wang P, Liu H, Luo J, Yu H, Huang J, Lin T: linc-UBC1 physically associates with polycomb repressive complex 2 (PRC2) and acts as a negative prognostic factor for lymph node metastasis and survival in bladder cancer. Biochimica et Biophysica Acta 2013, 1832:1528–1537.

Keene JD, Komisarow JM, Friedersdorf MB: RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nature protoc 2006, 1:302-7.

Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D: PRIDB: a Protein-RNA Interface Database. *Nucleic Acids Res* 2011, 39:D277-82.

Muppirala UK, Honavar V, Dobbs D: Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics* 2011, 12:489.

Ray D, Kazan H, Chan ET, Castillo LP, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR:. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nature Biotechnol 2009, 27:667-70.

Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Rattei T, Frishman D, Ruepp A: The Negatome database: a reference set of non-interacting protein pairs. Nucleic Acids Res 2010, 38:D540-4.

Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, Zhang Z, Skogerbø G, Chen L, Lu H, Zhao Y, Chen R: NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Res 2006, 34:D150-2.

APPENDIX A. IMPLEMENTATION OF RPISEQ AND RPINTDB

In this chapter, the back-end code and implementation of RPISeq webserver and RPIntDB database are provided.

The RPISeq webserver is developed using Perl, PHP and HTML. It has 4 different forms to accept input. The first form accepts only one protein and one RNA sequence at a time. This is the default page for RPISeq. 'batch-prot.html' allows user to upload a FASTA file of RNA sequences and a single protein sequence in plain text format. Similarly, 'batch_rna.html' accepts multiple protein sequences and a single RNA sequence. The pseudocode given below converts the input sequences into a feature vector as described in the original algorithm. Each line generated corresponds to a single RNA-pair.

```
#!/bin/perl

##NOTE:
## The model files used for predictions were built using Weka 3.7.0. The
models are incompatible with any other version of Weka.

## Input: [Protein file] [RNA file].
## The input files are in the usual FASTA format.

## HOW THE PROGRAM WORKS:

## 1. Load the protein sequences into a hash.
my %protein_sequences = ();

## 2. Load the RNA sequences into a hash
my %rna_sequences = ();

## The individual amino acids and nucleotides are separated into groups.
The mapping is given below.
my %proteinGroups = (
                'A' => 0,
                'G' => 0,
                'V' => 0,
                'I' => 1,
                'L' => 1,
```

```
                'F' => 1,
                'P' => 1,
                'Y' => 2,
                'M' => 2,
                'T' => 2,
                'S' => 2,
                'H' => 3,
                'N' => 3,
                'Q' => 3,
                'W' => 3,
                'R' => 4,
                'K' => 4,
                'D' => 5,
                'E' => 5,
                'C' => 6 );

my %rnaGroups = (
            'A' => 0,
            'U' => 1,
            'C' => 2,
            'G' => 3
          );

## Each line in the weka input file corresponds to a single protein-RNA
pair.

## We need to generate the input file for Weka prediction. The number of
input variables is equal to P^3 + R^4.
## The number P corresponds to the number of protein groups (7) and the
number R corresponds to the number of RNA groups (4).

## For the protein sequence, we count every 3-mer in the sequence (eg.
'GVI', 'LYC', etc.).

for (my $c = 0; $c < length (@protein) - 2; ++$c)
{
    my $three_mer_0 = @protein[$c];
    my $three_mer_1 = @protein[$c + 1];
    my $three_mer_2 = @protein[$c + 2];

    my $three_mer = $three_mer_0.$three_mer_1.$three_mer_2;

    ## Keep a count of the three-mers.
    $counts{$three_mer}++;
}

## Calculate the maximum and minimum counts.

$minimum = min(values %counts);
$maximum = max(values %counts);

## Calculate the weighted average of each 3-mer in the protein sequence
$weighted{$three_mer} = ($counts{$three_mer} - $minimum)/($maximum);
```

```
## For the RNA sequence, we do the same thing as above for every 4-mer in
the sequence (eg. 'AUUG', 'GCAC')



for (my $c = 0; $c < length(@rna) - 3; ++$c)
{
    my $four_mer_0 = @rna[0];
    my $four_mer_1 = @rna[$c + 1];
    my $four_mer_2 = @rna[$c + 2];
    my $four_mer_3 = @rna[$c + 3];

    my $four_mer = $four_mer_0.$four_mer_1.$four_mer_2.$four_mer_3;

    $counts{$four_mer}++;
}

$minimum = min(values %counts);
$maximum = max(values %counts);

$weighted{$four_mer} = ($counts{$four_mer} - $minimum)/($maximum);

## From these weighted counts, we can construct the input weka line
my $weka_line = "";
for (my $c = 0; $c < 343; ++$c)
{
    $weka_line .= $weighted{$three_mer};
}

for (my $c = 343; $c < 343 + 256; ++$c)
{
    $weka_line .= $weighted{$four_mer};
}
```

Add the arff header to the generated input. The arff header file lists the data types of the 599 features encoded in the vector. The first few lines and the last files of the arff header file are shown below.

```
@relation interactions

@attribute P1 NUMERIC
@attribute P2 NUMERIC
@attribute P3 NUMERIC
.
.
.
@attribute R254 NUMERIC
@attribute R255 NUMERIC
@attribute R256 NUMERIC
```

```
@attribute LABEL {1, 0}

@data
```

While making predictions on a single protein with multiple RNAs, the feature vector encoding of the protein is concatenated with every RNA encoding vector. For example, 1 protein and 20 RNAs will generate a 20-line weka input file. After appending the arff header to the beginning of the file, predictions are obtained by running the models on the weka input file.

The fourth form of RPISeq webserver is RPIntDB.html. Here, the user can submit a protein sequence to obtain homologous proteins in the database and their interacting RNA partners. The users have the option of adjusting the e-value for the BLAST run.

In RPIntDB, there are 3 tables: interaction, protein and rna. The schema for these tables are shown below.

The ‘interaction’ table contains the protein identifier and rna identifier of an interaction pair along with its source. Protein identifiers are typically UNIPROT identifiers. If there is a structure associated with the complex, the identifiers are PDB complex id with chain identifiers (e.g. 1ASY_A for protein and 1ASY_R for RNA).

Field	Type	Null	Key	Default	Extra
uid	int(11)	No	PRI	NULL	auto_increment
proteinid	varchar (60)	No		NULL	
rnaid	varchar (60)	No		NULL	
source	Text	No		NULL	

The ‘protein’ table contains information about the proteins listed in the ‘interaction’ table. The information includes name of the protein, sequence and PDB complex name (when structures are available). The ‘rna’ table contains similar information about the interacting RNAs.

Field	Type	Null	Key	Default	Extra
uid	int (11)	No	PRI	NULL	auto_increment
proteinid	varchar (60)	No		NULL	
Complex	Text	YES		NULL	
name	Text	YES		NULL	
sequence	Text	No		NULL	

Field	Type	Null	Key	Default	Extra
uid	Int (11)	No	PRI	NULL	auto_increment
rnaid	Varchar (60)	No		NULL	
Complex	Text	YES		NULL	
name	Text	YES		NULL	
sequence	Text			NULL	

All the protein sequences in the database are selected to create a FASTA file. ‘makeblastdb’ command is used to format the protein sequences for use with BLAST

program. When the user submits a protein sequence and specifies a threshold, the ‘blastp’ program is used to search the formatted protein database using the query sequence. If the user does not specify an e-value, a default value of 0.0001 is used. The BLAST results are then parsed to obtain the homologous proteins and the e-values for each hit. The protein hits are then used to query the ‘interaction’ table to obtain the interacting RNAs and the sources. The protein and RNA information for each pair are obtained from the ‘protein’ and ‘rna’ tables respectively.

Whenever RPIntDB is updated with new entries, new protein database has to be created to include all new protein sequences.

APPENDIX B. PRIDB V2.0: AN UPDATE TO THE PROTEIN-RNA INTERFACE DATABASE

Abstract

The Protein-RNA Interface Database (PRIDB) is a comprehensive database of protein-RNA interfaces extracted from protein-RNA complexes in the Protein Data Bank (PDB). It is designed to facilitate both detailed investigation of individual complexes and creation of custom datasets. PRIDB provides atomic- and residue-level interaction information for 1,484 protein-RNA complexes, comprising 16,350 protein chains and 3,398 RNA chains. Information about interactions and annotated motifs can be visualized within linear primary sequences of proteins or RNAs, and interfacial residues can be displayed in the context of three-dimensional structures, or in a machine-readable file format. Here, we present several new features of PRIDB: integration of RNA structural motifs from the RNA 3D Motif Atlas; refinement of the geometric rules used to define protein-RNA interactions; visualization of user-submitted structures, allowing detailed examination of structures not currently in the PDB; an additional non-redundant dataset, RB344, which includes annotations of protein-RNA complexes; and several performance improvements to increase user interface responsiveness and decrease computational time requirements. The PRIDB database is freely available at http://pridb.gdcb.iastate.edu.

Introduction

Protein-RNA interactions play important roles in many cellular and developmental processes. The results of the ENCODE Project (Djebali *et al.*, 2012) suggest that our understanding of the roles and prevalence of non-coding RNAs in the human genome remains largely incomplete. While high-throughput sequencing technology has led to exponential growth in the availability of RNA sequences, the corresponding growth in experimentally determined structure information has been considerably more modest, with protein-RAN complex structures comprising only ~1% of structures in the Protein Data Bank (PDB). Despite these limitations, careful analysis of detailed structural information has provided insights into fundamental principles of protein-RNA recognition (Borozan *et al.*, 2013) and characteristics of protein or RNA molecules involved in protein-RNA complexes (Iwakiri *et al.*, 2013, Ananth *et al.*, 2013). This information also informs computational methods, which have applied structural information to the problems of protein-RNA docking (Perez-Cano *et al.*, 2010, Huang *et al.*, 2013), protein-RNA interaction prediction (Puton *et al.*, 2012, Walia *et al.*, 2012), and protein-RNA partner prediction (Muppirala *et al.*, 2013, Cirillo *et al.*, 2013).

PRIDB is a repository of protein-RNA interfaces derived from structures in the Protein Data Bank (PDB). For each protein-RNA complex, PRIDB uses atomic coordinate information to calculate interface information using a distance threshold-based definition and geometric rule-based criteria. This information can be accessed as annotations on the primary sequence of each protein or RNA, as a three-dimensional display implemented via a JMol applet, or as a machine-readable CSV file. Users can also upload their own structures in PDB format and inspect them via any of these channels. In addition to interaction information,

PRIDB also integrates information about protein and RNA motifs from third-party sources (see below), with links to the original database. PRIDB's robust search function allows users to filter results by criteria such as experimental method, X-ray crystal resolution, protein or RNA length, or presence of a subsequence or motif. Several pre-calculated non-redundant datasets are also provided.

New features

Integration of RNA structural motifs from the RNA 3D Motif Atlas

Since the initial publication of PRIDB (Lewis *et al.*, 2011), a consistent nomenclature and accession scheme for structural motifs in RNA has been provided by the recently published RNA 3D Motif Atlas (http://rna.bgsu.edu/motifs) (Petrov *et al.*, 2013). The motifs from this resource are generated using FR3D (Sarver *et al.*, 2008), which was used by the previous version of PRIDB to annotate RNA structural motifs. PRIDB v2.0 has adopted the RNA 3D Motif Atlas accession scheme in its annotations.

Refinement of geometric interaction definitions

PRIDB calculates interacting residues in protein-RNA complexes using two different schemes: a distance-based definition, and a rule-based definition that considers the atomic geometries necessary for various types of physicochemical interaction. The first version of PRIDB used rules adapted from the program ENTANGLE (Allers and Shamoo, 2001); however, certain classes of contacts are not adequately differentiated by this rule set. Following the example of Treger and Westhof (2001), PRIDB v2.0 introduces two new classes of contacts: i) a 'clash' interaction, which represents close van der Waals contacts;

and ii) a ‘salt bridge’ interaction, which represents a hydrogen bond between a donor and acceptor that also form an electrostatic interaction. The definition of hydrogen bonding has also been updated to allow carbon to act as a hydrogen donor. A full list of geometric interaction definitions used by PRIDB v2.0 is presented in Table 1.

Visualization of user-submitted structures

PRIDB allows visualization of user-submitted structures in PDB format for interface calculation using either of the two interaction definitions described above. Whereas the first version of PRIDB returned that information in a machine-readable format only, PRIDB v2.0 also allows users to access user-submitted structures via the same interface used to view existing PDB structures in the database. This includes visualization of both annotated primary sequences and three-dimensional representations via a JMol applet.

Creation of a new non-redundant dataset, RB344

PRIDB provides several pre-calculated benchmark datasets for the convenience of users. These datasets are filtered to limit protein sequence redundancy and exclude low-resolution structures, making them ideal for use as input to computational methods that require protein-RNA interaction information as ‘training’ data. PRIDB v2.0 introduces an additional benchmark dataset, RB344, containing a total of 344 non-redundant protein chains and corresponding bound RNA chains. RB344 was calculated using the most recent data available in PRIDB and incorporates the modified geometric rules described above. In addition to interfacial information, RB344 is annotated to indicate the functional class (e.g., ribosomal, viral) of protein-RNA complexes in the dataset. The RB344 benchmark dataset

including both interface information and functional class annotations are available from the ‘Datasets’ section of the PRIDB homepage.

Performance enhancements and other improvements

The previous version of PRIDB used the BioPerl module (Stajich *et al.*, 2002) for all interface calculations. PRIDB v2.0 instead uses BioPython’s Bio.PDB module (Hamelryck and Manderick, 2003), which implements a KD tree in C++ to allow rapid lookup of atom-atom contacts. This reduces the computation time required for calculation of user-submitted complex interfaces and the time required to synchronize PRIDB with the PDB.

After the initial publication of PRIDB, we analyzed commonly used search criteria to guide the creation of additional database indexes; this has considerably improved the performance of SQL queries. This modification, coupled with other alterations to the front-end PHP code, has significantly increased user interface responsiveness during complex searches. Other portions of the user interface, such as the tutorial and FAQ section, have also been updated with the goal of improving usability.

Conclusions

In addition to the new features outlined above, the representation of protein-RNA complexes has grown substantially in the updated version of PRIDB. The number of protein-RNA complexes in PRIDB has increased from 926 (Lewis *et al.*, 2011) to 1,424 as of March 2013. Further, a 73% increase in the size of non-redundant datasets extracted from PRIDB (RB199 to RB344), reflects a significant increase in the diversity of protein-RNA complexes in the database. This richer database of interactions, together with new features, such as the

inclusion of RNA structural motifs from the RNA 3D Motif Atlas and improved rules that more finely differentiate classes of interactions, should make PRIDB v2.0 a valuable resource for researchers studying protein-RNA interactions.

Funding

This work is supported in part by funding from the National Institutes of Health [GM066387 to D.D. and V.H.]; and the National Science Foundation [DBI0923827 to D.D.].

Acknowledgements

We thank Xue Li and Pete Zaback for helpful discussions and suggestions.

References

Allers J, Shamoo Y: Structure-based analysis of protein-RNA interactions using the program ENTANGLE. J. Mol. Biol., 2001, 311, 75-86.

Ananth P, Goldsmith G, Yathindra N: An innate twist between Crick's wobble and Watson-Crick base pairs. RNA 2013, 19, 1038-1053.

Borozan SZ, Dimitrijević BP, Stojanović SD: Cation−π interactions in high resolution protein−RNA complex crystal structures. Comput Biol Chem 2013, 47, 105-112.

Cirillo D, Agostini F, Tartaglia GG: Predictions of protein-RNA interactions. Wiley Interdiscip Rev: Comput Mol Sci 2013, 3, 161–175.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al: Landscape of transcription in human cells. Nature 2012, 489, 101-108.

Hamelryck T, Manderick B: PDB file parser and structure class implemented in Python. Bioinformatics 2003, 19, 2308-10.

Huang Y, Liu S, Guo D, Li L, Xiao Y: A novel protocol for three-dimensional structure prediction of RNA-protein complexes. Sci Rep 2013, 3, 1887. doi: 10.1038/srep01887.

Iwakiri J, Kameda T, Asai K, Hamada M: Analysis of base-pairing probabilities of RNA molecules involved in protein-RNA interactions. Bioinformatics 2013, doi: 10.1093/bioinformatics/btt453.

Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D: PRIDB: A Protein–RNA Interface Database. Nucleic Acids Res. 2011, 39, D277-D282.

Muppirala UK, Lewis BA, Dobbs D: Computational tools for investigating RNA-protein interaction partners. J. Comput. Sci. Syst. Biol. 2013, 6, 182-187. doi:10.4172/jcsb.1000115.

Perez-Cano L, Solernou A, Pons C, Fernandez-Recio J: Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. Pac Symp Biocomput 2010, 293-301.

Petrov AI, Zirbel CL, Leontis NB: Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA 2013, 19, 1327-40.

Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM: Computational methods for prediction of protein-RNA interactions. J Struct Biol 2012, 179, 261–8.

Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB: FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. J. Math. Biol. 2008, 56, 215-252.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H. et al: The Bioperl toolkit: Perl modules for the life sciences. Genome. Res. 2002, 12, 1611-8.

Treger M, Westhof E: Statistical analysis of atomic contacts at RNA-protein interfaces. J. Mol. Recognit. 2001, 14, 199-214.

Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, El-Manzalawy Y, Dobbs D, Honavar V: Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinformatics 2012, 13, 89.

ACKNOWLEDGEMENTS

My advisors, colleagues, friends and family have helped me in so many ways to successfully complete this dissertation. I express my thanks to all who have contributed to my graduate education.

First of all, I would like to thank my advisor Drena Dobbs. She has been very supportive and helpful throughout my graduate career. She provided me the opportunity to work on research problems that are of interest to me. She is the most helpful, generous and kind teacher I have ever had the good fortune to encounter. I would also like to thank my co-major professor Robert Jernigan, whose guidance has been influential in this dissertation. I would also like to thank Heike Hofmann, Edward Yu and Guang Song for serving on my committee and for their helpful discussions and suggestions.

I would like to thank my parents who have always encouraged me to aim higher and strive towards my goals. I will be forever grateful for their support. If not for my mom, I would never have had the opportunity to pursue my studies in the US. A special thanks to my sister Silpha, who has been there for me in painful times and helped me focus completely on my studies. I am thankful for Muruganath, my friend, philosopher and guide, who has helped me in every step of my education. I would also like to thank all of my family members and friends for their encouragement and support.

I have been very fortunate to work with Ben Lewis, Rasna Walia, Deepak Reyon and Pete Zaback in the Dobbs lab. It has been a pleasure to collaborate with Ben on research projects. Thanks to Ben, I have submitted manuscripts, abstracts and my final dissertation on time. I consider myself lucky to be a part of the Dobbs lab where every day is fun and

productive. I am thankful for the long discussions with them that led to valuable insights in new research problems.

Most of all, I would like to thank my husband Mridul, without whom I'd have been lost. You have been very patient, supportive and encouraging all these years. Because of you, Mridul, my graduate life has been an awesome experience.